Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Σηματων, Ελεγχου και Ρομποτικης

Προσαρμοστική Ενισχυτική Μηχανική Μάθηση για την ανάπτυξη Ρομποτικών Δεξιοτήτων σε Δυναμικά Περιβάλλοντα

Διπλωματικη Εργασια

του

ΓΕΩΡΓΙΟΥ ΒΕΛΕΝΤΖΑ

Επιβλέπων: Κωνσταντίνος Τζαφέστας
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εργαστηριο Ευφυων Ρομποτικων Συστηματων και Αυτοματισμου

Αθήνα, Οκτώβριος 2018

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Ευφυών Ρομποτικών Συστημάτων και Αυτοματισμού

Προσαρμοστική Ενισχυτική Μηχανική Μάθηση για την ανάπτυξη Ρομποτικών Δεξιοτήτων σε Δυναμικά Περιβάλλοντα

Διπλωματικη Εργασια

του

ΓΕΩΡΓΙΟΥ ΒΕΛΕΝΤΖΑ

Επιβλέπων: Κωνσταντίνος Τζαφέστας
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 22η Οκτωβρίου 2018.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Κωνσταντίνος Τζαφέστας	Πέτρος Μαραγκός	Γεώργιος Στάμου
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π

Αθήνα, Οκτώβριος 2018

(Υπογραφή)

.....

Γεωργιος Βελεντζας

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη νέων μεθόδων προσαρμοστικής ενισχυτικής μηχανικής μάθησης με κύριο πεδίο εφαρμογής την αλληλεπίδραση ανθρώπου-ρομπότ. Η έρευνα αυτή ξεκινάει από το θεμελιώδες πρόβλημα της στοχαστικής βελτιστοποίησης αποφάσεων σε μία κατάσταση, ένα πρόβλημα που είναι γνωστό στην παγκόσμια βιβλιογραφία ως multi-armed bandit task, ενώ θα αποδοθεί στην ελληνική γλώσσα ως μηχανή επιβράβευσης πολλαπλών επιλογών. Στη συνέχεια οι ιδέες αυτές επεκτείνονται σε χρονομεταβλητές μαρκοβιανές διαδικασίες λήψης αποφάσεων άγνωστης δομής, προσεγγίζοντας το δίλημμα εξερεύνησης-αξιοποίησης (exploration-exploitation) με τεχνικές εμπνευσμένες από τον τομέα των νευροεπιστημών.

Το πρόβλημα εκτίμησης της βέλτιστης (δυναμικά εξελισσόμενης) αναλογίας εξερεύνησης-αξιοποίησης έχει μελετηθεί εκτενώς στη βιβλιογραφία από τα πεδία Μηχανικής Μάθησης και Υπολογιστικής Νευροεπιστήμης. Στην εργασία αυτή παρουσιάζεται αρχικά μία προσπάθεια για γεφύρωση των δύο κλάδων με την ανάπτυξη ενός υβριδικού αλγορίθμου, συνδυάζοντας βιολογικά εμπνευσμένη μετα-μάθηση με φίλτρα Kalman και επιβραβεύσεις εξερεύνησης. Συγκρίνοντας την επιτευχθείσα απόδοση με αυτή σύγχρονων και επίκαιρων δυναμικών αλγορίθμων σε ένα σύνολο αριθμητικών προσομοιώσεων διαφορετικών σεναρίων, ο υβριδικός αλγόριθμος φαίνεται να συνδυάζει τα πλεονεκτήματα των μεθόδων και επιδεικνύει καλύτερη συμπεριφορά των προγενέστερων.

Στη συνέχεια, προτείνεται ένας προσαρμοστικός αλγόριθμος ενισχυτικής μάθησης με παραμετροποιημένες διακριτές δράσεις και εμπλουτισμένος με στρατηγική ενεργής εξερεύνησης ανά κατάσταση. Η εφαρμοσιμότητά του επιδεικνύεται σε κλασσικά προβλήματα, όπως αυτό της πλοήγησης σε άγνωστο χάρτη, καθώς και με την βελτιστοποίηση της αλληλεπίδρασης ρομπότ-παιδιού παράλληλα με την εκμάθηση επίλυσης του παζλ «ο πύργος του Ανόι».

Λέξεις Κλειδιά

Ενισχυτική Μηχανική μάθηση, Μαρκοβιανές διαδικασίες λήψης αποφάσεων, Μηχανές επιβράβευσης πολλαπλών επιλογών, Φίλτρο Kalman, Προσαρμοστικότητα, Δίλημμα εξερεύνησης-αξιοποίησης, Αλληλεπίδραση ανθρώπου-ρομπότ.

Abstract

The purpose of this diploma thesis is to develop new approaches and methods of adaptive reinforcement learning which will be mainly implemented on human-robot interaction scenarios. This research starts from the fundamental problem of stochastic optimization of decision making in one single state, a problem in the literature which is well known as a multi-armed bandit task. The ideas are then expanded on non-stationary Markov decision processes of an unknown structure, tackling the exploration-exploitation dilemma with a bio-inspired method from the field of computational neuroscience.

The problem of finding an efficient (dynamically changing) exploration-exploitation trade-off has been well studied both in the Machine Learning and Computational Neuroscience fields. The first objective of this work is to bridge some of the different methods of these two fields by implementing a hybrid algorithm which combines bio-inspired meta-learning, Kalman filter, and exploration bonuses. The performance of the algorithm is then compared to several state-of-the-art alternatives on a set of non-stationary stochastic multi-armed bandit tasks, where it displays a good combination of advantages from different methods and outperforms these methods in the studied scenarios.

The ideas are then expanded in multi-state dynamically changing environments by developing an adaptive reinforcement learning algorithm with parameterized actions and state-specific exploration. Its applicability and adaptive nature is then demonstrated on a number of problem sets, like a continuous maze problem as an enhancement of the classic grid world which is used as a benchmark in artificial intelligence and robotics, as well as in a simulated human-robot interaction where the robot's objective is to maximize a child's engagement/attention while learning to solve the known puzzle ≪tower of Anoi≫.

Keywords

Reinforcement learning, Markov decision processes, Multi-armed bandits, Kalman Filter, Adaptivity, Exploration-exploitation trade-off, Human-robot interaction

Ευχαριστίες

Ευχαριστώ τον καθηγητή μου κ. Κωνσταντίνο Τζαφέστα αφενώς για την εξαίρετη διδασκαλία των μαθημάτων του που αποτέλεσε το εφαλτήριο για την ανάπτυξη του ερευνητικού μου ενδιαφέροντος στον τομέα της Ρομποτικής, και αφετέρου για τη δυνατότητα που μου έδωσε μέσα από μια ερευνητική διπλωματική εργασία να έρθω σε επαφή με την επιστημονική κοινότητα σε βαθμό υπεράνω των προσδοκιών μου. Τον ευχαριστώ ιδιαίτερα για την εμπιστοσύνη που έδειξε στο πρόσωπό μου με τη συμμετοχή μου στο ευρωπαϊκό έργο BabyRobot, Horizon2020, κατα την οποία είχα τη δυνατότητα να παρευρίσκομαι και να συμμετέχω σε συναντήσεις υψηλού ερευνητικού επιπέδου ως μέλος του εργαστηρίου Ρομποτικής και Αυτοματισμού και του Επιστημονικού Ινστιτούτου Συστημάτων Επικοινωνιών και Υπολογιστών. Παράλληλα ευχαριστώ τον Dr. Mehdi Khamassi, Research Scientist, Centre National de la Recherche Scientifique, για τις ιδέες που μοιράστηκε μαζί μου καθώς και για την συνολική συνεπίβλεψη και καθοδήγηση της διπλωματικής μου που εντέλει οδήγησε στην συγγραφή και συν-συγγραφή συνολικά τέσσάρων επιστημονικών δημοσιεύσεων διεθνών συνεδρίων και δύο δημοσιεύσεων σε διεθνή έγκριτα περιοδικά. Ευχαριστώ και τους δύο για την υποστήριξη αλλά και για την εμπιστοσύνη που μου έδειξαν ώστε να παρουσιάσω το ερευνητικό μας έργο σε τρία από αυτά τα συνέδρια.

Επίσης, θα ήθελα να ευχαριστήσω τον καθηγητή κ. Πέτρο Μαραγκό καθώς τα μαθήματα Ψηφιακής Επεξεργασίας Σημάτων, Όρασης Υπολογιστών, Αναγνώρισης Προτύπων αλλα και το μεταπτυχιακό μάθημα Θεωρητικών Μεθόδων Όρασης Υπολογιστών και Επεξεργασίας Σημάτων το οποίο μου έδωσε τη δυνατότητα να παρακολουθήσω, αποτέλεσαν για μένα σημαντική ακαδημαϊκή έμπνευση. Ευχαριστώ τον καθηγητή κ. Γεώργιο Στάμου για τον χρόνο του και την συμμετοχή του στην εξεταστική επιτροπή, αλλά και για τον ουσιαστικό τρόπο διδασκαλίας του στο μάθημα της Τεχνητής Νοημοσύνης. Ευχαριστώ επίσης τους καθηγητές κ. Νικόλαο Μαράτο, κ. Γεώργιο Φικιώρη και κ. Ηλία Κατσούφη για τις ενδιαφέρουσες συζητήσεις που είχαμε στη διάρκεια των διαλέξεων των μαθημάτων τους. Τέλος, ευχαριστώ την οικογένειά μου, και ιδιαίτερα τη μητέρα μου.

Γιώργος Βελεντζάς

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Ρομποτική και Ενισχυτική Μηχανική Μάθηση

Βρισκόμενοι στην εποχή των μεγάλων δεδομένων και της πληροφορίας, μία εκ των μεγαλύτερων προκλήσεων είναι η προσεκτικά δομημένη μοντελοποίηση της διαδικασίας της μάθησης με τέτοιο τρόπο ώστε καινοτόμα αυτόνομα συστήματα να εκμεταλλεύονται την προϋπάρχουσα γνώση και να δημιουργούν νέα εδάφη αναπτύσσοντας χαρακτηριστικά που θα προσεγγίζουν τις ανθρώπινες γνωσιακές ικανότητες. Οι ήδη υπάρχουσες προσεγγίσεις μέσω των τριών πυλώνων της μηχανικής μάθησης, της μάθησης με επίβλεψη, της μάθησης χωρίς επίβλεψη, και της ενισχυτικής μάθησης, παρέχουν μία πληθώρα εντυπωσιακά δομημένων μαθηματικών μεθόδων, δίνοντας παράλληλα στην ερευνητική κοινότητα μια ισχυρή εργαλειοθήκη για πειραματισμό και ανάπτυξη χρήσιμων κοινωνικών εφαρμογών.

Αυτόματος Έλεγχος και Μηχανική Μάθηση

Αν και οι κλασσικές προσεγγίσεις μέσω αυτομάτου ελέγχου, παρέχουν τις απαραίτητες εγγυήσεις και προδιαγραφές ανάλογα με την εφαρμογή, το πλαίσιο της ενισχυτικής μάθησης παρέχει τη δυνατότητα για προσέγγιση προβλημάτων που δεν είναι εύκολο να μοντελοποιηθούν. Χαρακτηριστικά, όπως αναφέρουν οι Kober, Bagnell, Peters στο [31],

> *"Η ενισχυτική μάθηση προσφέρει στον τομέα της ρομποτικής το κατάλληλο πλαίσιο και τα εργαλεία για τη σχεδίαση δύσκολα μοντελοποιήσιμων συμπεριφορών. Από την άλλη πλευρά, η ρομποτική προσφέρει στον τομέα της ενισχυτικής μάθησης την έμπνευση για ανάπτυξη νέων μεθόδων που θα έχουν ισχυρό αντίκτυπο σε ένα ενθουσιώδες πεδίο εφαρμογών..."*

Φυσικά το δίλημμα και η διαφωνία μεταξύ μίας μαθηματικά εγγυημένης προσέγγισης ή μίας εμπειρικής προσέγγισης δεν είναι κάτι νέο. Ως παράδειγμα, οι Αρχαίοι Έλληνες προτιμούσαν τη δημιουργία μοντέλων για την εξήγηση και πρόβλεψη φαινομένων, ενώ οι Βαβυλώνιοι προτιμούσαν να κάνουν προβλέψεις βάσει εμπειρικών προσεγγίσεων, με εξίσου καλά αποτελέσματα. Έτσι και σήμερα, η δημιουργία ενός εμπειρικού αλγορίθμου μάθησης για την επιτυχή σταθεροποίηση του ανάστροφου εκκρεμούς μπορεί να έχει το ίδιο καλά αποτελέσματα με την κλασσική

προσέγγιση της θεωρίας ελέγχου. Η απάντηση όμως στο ερώτημα για το ποια μέθοδος είναι καλύτερη και ποια θα προτιμούσε κανείς ανάλογα με την εφαρμογή, δεν είναι εύκολο να απαντηθεί. Αντί αυτού θα επισημάνουμε την ομοιότητα των εξισώσεων βέλτιστου ελέγχου Hamilton-Jacobi-Bellman με τις εξισώσεις ανανέωσης των συναρτήσεων αξίας δράσεων και καταστάσεων στο πλαίσιο της ενισχυτικής μάθησης, κάτι που μπορεί να δώσει κίνητρα στον αναγνώστη για περαιτέρω διερεύνηση.

Στις περιπτώσεις όπου οι μαθηματικές εξισώσεις που διέπουν το σύστημα είναι γνωστές (ή κατά προσέγγιση γνωστές) και οι παράμετροι του προβλήματος (π.χ. η ροπή αδράνειας του συνδέσμου και η επιτάχυνση της βαρύτητας στο μοντέλο του ανάστροφου εκκρεμούς) είναι δεδομένες, τότε σαφώς η θεωρία ελέγχου δίνει περισσότερες εγγυήσεις. Ακόμα και στην περίπτωση που έχουμε μία απλή εκτίμηση των παραμέτρων, μέθοδοι προσαρμοστικού ελέγχου με μοντέλα αναφοράς (model reference adaptive control) [45, 36], ή μέθοδοι άπειρου ορίζοντα (infinite-horizon model predictive control) [10] μπορούν να εφαρμοστούν κατάλληλα. Σε προβλήματα όμως που δεν υπάρχει αρκετή γνώση για την μοντελοποίηση των σχέσεων εισόδων-εξόδων, ενώ παράλληλα δεν υπάρχει καμία εγγύηση για την παρατηρησιμότητα και την ελεγξιμότητα του συστήματος, τότε η θεωρία ελέγχου δεν μπορεί να εφαρμοστεί, ενώ οι μέθοδοι ενισχυτικής μάθησης φαίνεται να δίνουν υποσχόμενα εμπειρικά αποτελέσματα.

1.2 Κίνητρα και Προκλήσεις

Συνεχίζοντας τις ιδέες από την προηγούμενη συζήτηση και έχοντας ως παράδειγμα ένα περιβάλλον αλληλεπίδρασης ανθρώπου-ρομπότ, είναι σαφές ότι δεν μπορούμε με ασφαλή τρόπο να μοντελοποιήσουμε τη συμπεριφορά του ανθρώπου ως προς τον τρόπο που αλληλεπιδρά, είτε με λεκτικά σήματα, είτε με μη-λεκτικά σήματα. Εάν θεωρήσουμε ότι έχουμε πλήρη έλεγχο των αρθρώσεων ενός μικρού σύγχρονου ανθρωποειδούς ρομπότ, ενισχυμένο με τη δυνατότητα 3Δ ανακατασκευής του περιβάλλοντος, αναγνώρισης αντικειμένων, και άλλου τύπου οπτικοακουστικές δεξιότητες, και θεωρήσουμε ως έξοδο του συστήματος χαρακτηριστικά που αφορούν την ανθρώπινη συμπεριφορά, από τα απλούστερα όπως τη θέση του ανθρώπου στον χώρο, έως πιο σύνθετα χαρακτηριστικά που αφορούν την ψυχοσυναισθηματική κατάσταση του ανθρώπου όπως τη διάθεση (χαρά, λύπη, ενθουσιασμό), τότε ποιο είναι το μοντέλο ελέγχου; Μπορούμε να ελέγξουμε τα συναισθήματα ή τα κίνητρα; Άν και τέτοιου είδους ερωτήματα χρήζουν ιδιαίτερης βιοηθικής αντιμετώπισης δεν παύουν να μας απασχολούν καθημερινά, όταν για παράδειγμα στη θέση του μικρού ρομπότ βρισκόμαστε εμείς και στη θέση του ανθρώπου ένας κακοδιάθετος φίλος.

1.2.1 Αλληλεπίδραση Ανθρώπου-Ρομπότ

Μεταξύ του συνόλου των ρομποτικών εφαρμογών που υπάρχουν, ή που βρίσκονται προς ανάπτυξη, πολλές αφορούν τα ρομπότ υποβοήθησης. Σε αυτές τις εφαρμογές υπάρχει συνήθως και η ανάγκη για αναγνώριση μη-λεκτικών χαρακτηριστικών από το ρομπότ κατά την αλληλεπίδρασή του με τον άνθρωπο, και η επιπλέον ανάγκη για ικανότητα δυναμικής προσαρμοστικότητας [17]. Ως παράδειγμα, ένα ρομπότ που υποβοηθά ηλικιωμένους ανθρώπους στο

σπίτι, θα πρέπει να μπορεί να ανιχνεύσει πότε ο άνθρωπος έχει δυσκολία στο να πιάσει ένα αντικείμενο, να σηκωθεί, ή να περπατήσει, ώστε να αντιδράσει άμεσα για την παροχή βοήθειας. Ένα άλλο παράδειγμα σχετίζεται με εκπαιδευτικές εφαρμογές, όπου ένα μικρό ανθρωποειδές ρομπότ μπορεί να υποβοηθήσει τον άνθρωπο-εκπαιδευτικό ώστε να ενισχύσει το ενδιαφέρον παιδιών υπό το φάσμα του αυτισμού (ASD) σε εκπαιδευτικά παιχνίδια, να τα βοηθήσει για την ανάπτυξη κοινωνικών δεξιοτήτων [26, 48, 5] και την ενίσχυση των κινήτρων τους [11].

Στις περιπτώσεις ρομπότ υποβοήθησης για εκπαιδευτικές εφαρμογές, ένα σύνηθες μηλεκτικό σήμα για το οποίο φαίνεται να υπάρχει μεγάλο ενδιαφέρον από την επιστημονική κοινότητα, αφορά την αύξηση της συμμετοχής/προσοχής[1] του παιδιού κατά τη διάρκεια κάποιας εργασίας [47, 25, 38, 51]. Υπάρχει μια πληθώρα από διαφορετικές μετρικές που έχουν αναπτυχθεί γι΄ αυτόν τον σκοπό, πολλές εκ των οποίων βασίζονται σε χαρακτηριστικά της στάσης του σώματος και της διεύθυνσης του βλέμματος [47, 1]. Συνολικά βέβαια, η πρόταση ενός πλήρους μοντέλου για εκτίμηση της συμμετοχής/προσοχής είναι πολύπλοκη και δεν είναι αντικείμενο μελέτης της συγκεκριμένης εργασίας. Ο σκοπός της εργασίας αυτής αφορά εφαρμογές βελτίωσης της ικανότητας του ρομπότ στο να προσαρμόζει γρήγορα τη συμπεριφορά του μετά από παρατήρηση αλλαγής της στάσης ή του βλέμματος του παιδιού, έτσι ώστε να βελτιώσει την αποδοτικότητα και τους στόχους ενός εκπαιδευτικού παιχνιδιού.

Οι προοπτικές για ανάπτυξη ρομποτικών δεξιοτήτων σε αυτού του τύπου κοινωνικές εφαρμογές καθορίζονται από προκλήσεις που εμφανίζονται και σε άλλες εφαρμογές, μη κοινωνικού τύπου. Κατά πρώτον, πολλά προβλήματα μπορούν να μοντελοποιηθούν είτε χρησιμοποιώντας ένα διακριτό σύνολο γενικευμένων δράσεων (π.χ. πιάσε, δείξε), είτε σε έναν συνεχή χώρο δράσεων (π.χ. άμεσος έλεγχος των αρθρώσεων). Κατά δεύτερον ένα σημαντικό ζήτημα προς διερεύνηση είναι η κατάλληλη επιλογή του επιπέδου εξερευνητικών δράσεων κατά τη διαδικασία μάθησης ώστε το ρομπότ να είναι ικανό να αντεπεξέλθει σε πιθανή μη στατικότητα κατά τη διάρκεια κοινωνικών αλληλεπιδράσεων.

Αρκετές έρευνες έχουν ήδη εφαρμόσει τεχνικές ενισχυτικής μάθησης με χρήση διακριτού χώρου δράσεων, συμπεριλαμβανομένων εφαρμογών αλληλεπίδρασης ανθρώπου-ρομπότ (π.χ [30]). Παρ΄ όλα αυτά η απλοποίηση ενός γενικού προβλήματος με διακριτοποίηση των διαθέσιμων δράσεων σε ένα μικρό πεπερασμένο σύνολο, προϋποθέτει πολλές φορές τον παράγοντα ανθρώπινης τεχνογνωσίας και ίσως δεν είναι εφικτή σε πιο σύνθετα προβλήματα στα οποία ο άμεσος συνεχής έλεγχος των ρομποτικών αρθρώσεων είναι η μόνη επιλογή. Οι εφαρμογές ενισχυτικής μάθησης σε συνεχείς χώρους δράσης [32, 55] είναι επίσης πολλά υποσχόμενες για την επίτευξη προσαρμοστικής συμπεριφοράς των ρομπότ σε πραγματικά περιβάλλοντα αλληλεπίδρασης [31]. Ωστόσο είναι και πάλι απαραίτητος ο ανθρώπινος παράγοντας κατά τη διαδικασία μάθησης (π.χ μάθηση με επίδειξη), ώστε να καθοδηγήσει την αναζήτηση της πολιτικής σε ένα μικρό υποσύνολο του χώρου δράσεων, το αρχικό μέγεθος του οποίου είναι απαγορευτικό.

[1] Ως συμμετοχή εννοούμε την διάθεση για αλληλεπίδραση, προσοχή και συνολικό ενδιαφέρον. Στην παγκόσμια βιβλιογραφία είναι γνωστή με τον αγγλικό όρο engagement.

1.2.2 Γενικό Πειραματικό Παράδειγμα

Το γενικό πειραματικό παράδειγμα που υιοθετείται σε αυτή την εργασία, αφορά την αλληλεπίδραση ενός μικρού ανθρωποειδούς ρομπότ (π.χ NAO, Zeno) με ένα παιδί (υπό την επίβλεψη ενός ενήλικα), όπου ο σκοπός θα είναι η μεγιστοποίηση της συμμετοχής του παιδιού σε κάποια εργασία. Το παράδειγμα αυτό καθορίζεται από το πλαίσιο του χρηματοδοτούμενου προγράμματος Ευρωπαϊκής Ένωσης BabyRobot (H2020-ICT-24-2015-6878310), κατά το οποίο σχεδιάστηκε ένα σύνολο από εφαρμόσιμα σενάρια αλληλεπίδρασης για τη μελέτη και την ανάπτυξη ειδικών κοινωνικό-συναισθηματικών και επικοινωνιακών χαρακτηριστικών σε παιδιά υπό το φάσμα το αυτισμού αλλά και τυπικά αναπτυσσόμενα παιδιά (TD). Σε αυτό το πλαίσιο, σχεδιάστηκε μία πιλοτική πειραματική διάταξη στην οποία ένα ρομπότ NAO αλληλεπιδρά με ένα παιδί και επαναλαμβανόμενα εκφράζει ενδιαφέρον στο να πιάσει κάποιο αντικείμενο (π.χ έναν μικρό κύβο) το οποίο όμως δεν είναι βρίσκεται στον χώρο δράσης του. Το ρομπότ δείχνει το αντικείμενο στο παιδί αλλάζοντας την εκφραστικότητα της κίνησης (π.χ ανοίγοντας και κλείνοντας την παλάμη με κάποια ταχύτητα, μετακινώντας τον κορμό του με κλίση προς το αντικείμενο με κάποια γωνία, κοιτάζοντας το παιδί με διαφορετικά περιοδικά διαστήματα) έως ότου το παιδί αντιληφθεί την πρόθεση του ρομπότ, εκφράσει ενδιαφέρον και βοηθήσει το ρομπότ να πιάσει το αντικείμενο. Υπό την υπόθεση ότι κάθε παιδί έχει διαφορετικές προτιμήσεις στα επίπεδα εκφραστικότητας κατά την επικοινωνία, είναι επιθυμητό να μπορεί το ρομπότ να προσαρμοστεί αλλάζοντας τα επίπεδα εκφραστικότητας μέσω των παραμέτρων που περιγράφουν τις κινήσεις του. Επιπρόσθετα, οι προτιμήσεις ενός παιδιού μπορεί να είναι μη-στατικές, και να αλλάζουν είτε κατά τη διάρκεια μίας συνεδρίας αλληλεπίδρασης είτε σε μεγαλύτερο εύρος χρόνου. Αυτό θα συνεπάγεται μείωση του ενδιαφέροντος και της συμμετοχής, και γι΄ αυτό είναι απαραίτητη η ανάπτυξη προσαρμοστικής εκφραστικότητας σε τέτοια περιβάλλοντα. Σε ένα δεύτερο πιλοτικό σενάριο, το ρομπότ και το παιδί προσπαθούν να επιλύσουν ένα παζλ (όπως τον πύργο του Ανόι). Σε αυτή την περίπτωση το πρόβλημα περιγράφεται από διαφορετικές καταστάσεις, όπου η κάθε κατάσταση μπορεί να αφορά τις θέσεις των κύβων (ή τόρων στο κλασσικό πρόβλημα). Ο στόχος σε αυτό το σενάριο μπορεί να αφορά την ανάπτυξη κοινωνικού τύπου δεξιοτήτων, όπου ιδανικά το παιδί θα συμμετέχει στην επίλυση του προβλήματος όταν το ρομπότ δεν φαίνεται να τα καταφέρνει.

1.3 Αντικείμενο και Στόχοι της Εργασίας

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη νέων μεθόδων προσαρμοστικής ενισχυτικής μηχανικής μάθησης με κύριο πεδίο εφαρμογής την αλληλεπίδραση ανθρώπου-ρομπότ. Η έρευνα αυτή ξεκινάει από το θεμελιώδες πρόβλημα της στοχαστικής βελτιστοποίησης αποφάσεων σε μία κατάσταση, ένα πρόβλημα που είναι γνωστό ως στην παγκόσμια βιβλιογραφία ως multi-armed bandit task (MAB). Για την επίλυση του προβλήματος λήψης αποφάσεων υπό αβεβαιότητα, είναι απαραίτητη η ρύθμιση μίας δυναμικά εξελισσόμενης αναλογίας μεταξύ εξερευνητικών και αξιοποιητικών δράσεων. Στα πεδία της ρομποτικής μάθησης [31, 28, 6] και των υπολογιστικών νευροεπιστημών μάθησης και λήψης αποφάσεων

[15, 12, 27, 58], αρκετές έρευνες αντιμετωπίζουν το δίλημμα εξερεύνησης-αξιοποίησης με τη χρήση της συνάρτησης soft-max Boltzmann, χρησιμοποιώντας επιβραβεύσεις εξερεύνησης για κάποιες δράσεις (exploration bonuses) [13, 18]. Παρόλο που η προσέγγιση αυτή είναι πολλά υποσχόμενη λόγω των πολύ καλών εμπειρικών αποτελεσμάτων, υπάρχουν αρκετοί περιορισμοί που μπορεί να σκεφτεί κανείς όταν το περιβάλλον είναι μη στατικό. Στην εργασία αυτή παρουσιάζεται αρχικά μία προσπάθεια για γεφύρωση των δύο κλάδων με την ανάπτυξη ενός υβριδικού αλγορίθμου, συνδυάζοντας βιολογικά εμπνευσμένη μετα-μάθηση με φίλτρα Kalman και επιβραβεύσεις εξερεύνησης. Συγκρίνοντας την επιτευχθείσα απόδοση με αυτή σύγχρονων και επίκαιρων δυναμικών αλγορίθμων σε ένα σύνολο αριθμητικών προσομοιώσεων διαφορετικών σεναρίων, ο υβριδικός αλγόριθμος φαίνεται να συνδυάζει τα πλεονεκτήματα των μεθόδων και επιδεικνύει καλύτερη συμπεριφορά των προγενέστερων.

Στη συνέχεια οι ιδέες αυτές επεκτείνονται σε χρονομεταβλητές μαρκοβιανές διαδικασίες λήψης αποφάσεων άγνωστης δομής. Στα [30, 29], προτείναμε ένα πλαίσιο των Μαρκοβιανών Διαδικασιών Λήψης Αποφάσεων σε Παραμετροποιημένους Χώρους Δράσεων (PAMDPs) [41, 23], και τη χρήση τους σε σενάρια αλληλεπίδρασης ανθρώπου-ρομπότ, καθώς αποτελούν μία υποσχόμενη μέση λύση μεταξύ των διακριτών και των συνεχών χώρων δράσης. Το πλαίσιο αυτό επιτρέπει την ανάπτυξη ενός πλούσιου συμπεριφορικού ρεπερτορίου του ρομπότ, ώστε να μπορεί να επιλέξει μεταξύ ενός μικρού συνόλου διακριτών δράσεων, όπως το να κλωτσήσει μία μπάλα, να στρίψει, ή να τρέξει, ενώ παράλληλα να μάθει τις κατάλληλες παραμέτρους για κάθε δράση, όπως τη δύναμη του λακτίσματος, τη γωνία στροφής ή την ταχύτητα βάδισης αντίστοιχα. Επιπλέον, για την εφαρμογή αυτού του πλαισίου σε μη-στατικά προβλήματα, εισάγαμε έννοιες ενεργούς αναζήτησης [49, 4, 43, 7, 64] και αποφύγαμε τη χρήση ενός σταθερού κλάσματος μεταξύ εξερευνητικών και αξιοποιητικών δράσεων. Αρχικά επικεντρωθήκαμε περισσότερο σε σενάρια μίας κατάστασης, κάτι που διερευνήσαμε επιπρόσθετα στα [61, 62]. Αυτό μας επέτρεψε να προβούμε στην αναζήτηση εμπειρικών αποδεικτικών στοιχείων για την αποδοτική συμπεριφορά των αλγορίθμων μας, αναλύοντας χαρακτηριστικά ευρωστίας, αβεβαιότητας και διακυμάνσεων της ανθρώπινης συμμετοχής κατά τη διάρκεια της αλληλεπίδρασης [30]. Σε αυτή την εργασία το παραπάνω πλαίσιο επεκτείνεται και σε περισσότερες καταστάσεις ώστε να επιτύχουμε τη γενίκευσή του σε μία ποικιλία εκπαιδευτικών παιχνιδιών που μπορούν να χρησιμοποιηθούν σε περιβάλλοντα αλληλεπίδρασης ρομπότ-παιδιού. Μία επίσης σημαντική καινοτομία αυτής της επέκτασης είναι η πρόταση για *εξειδικευμένη εξερεύνηση ανά κατάσταση*, η οποία επιτρέπει καλύτερη διαχείριση όταν η μη στατικότητα αφορά μόνο ένα υποσύνολο του χώρου κατάστασης, ενώ κάποια τμήματα του χώρου κατάστασης είναι στατικά. Επιπρόσθετα, η κάθε δράση μπορεί να περιγραφεί με ένα πλήθος παραμέτρων αντί μίας, οι οποίες θα πρέπει να προσαρμόζονται κατάλληλα σε πραγματικό χρόνο, συνεπώς επεκτείναμε την ικανότητα του αλγορίθμου προς αυτό το σκοπό.

1.3.1 Συνεισφορά

Στα πλαίσια της εργασίας υπήρχε μία πληθώρα από ερευνητικές δραστηριότητες σε συνεργασία με το Εργαστήριο Ρομποτικής και Αυτομάτου ελέγχου του Εθνικού Μετσόβιου

Πολυτεχνείου και το Επιστημονικό Ινστιτούτο Συστημάτων Επικοινωνιών και Υπολογιστών. Η συνεισφορά γενικού σκοπού της εργασίας αφορά:

• την ανάπτυξη ενός υβριδικού προσαρμοστικού αλγορίθμου για λήψη αποφάσεων υπό αβεβαιότητα σε δυναμικά περιβάλλοντα μίας κατάστασης.

• την ανάπτυξη ενός προσαρμοστικού αλγορίθμου ενισχυτικής μάθησης με εξειδικευμένη εξερεύνηση ανά κατάσταση για παραμετροποιημένους χώρους δράσεων.

ενώ οι ερευνητικές δραστηριότητες που έγιναν στα πλαίσια της εργασίας και αφορούν την παρουσίαση του επιστημονικού έργου της ερευνητικής ομάδας είναι:

• παρουσίαση της δημοσίευσης με τίτλο Active exploration and parameterized reinforcement learning applied to a simulated human-robot interaction task, ως συν-συγγραφέας στο International Robotic Computing Conference 2017, Taichung, Taiwan.

• παρουσίαση με poster της εργασίας με τίτλο Bridging Computational Neuroscience and Machine Learning on Non-Stationary Multi-Armed Bandits, στο Reinforcement Learning and Decision Making Conference 2017, Ann Arbor, USA.

• παρουσίαση της δημοσίευσης με τίτλο Bio-inspired meta-learning for active exploration during non-stationary multi-armed bandit tasks, ως κύριος συγγραφέας στο συνέδριο Intelligent Systems Conference 2017, London, UK.

Σημαντικό μέρος των μεθόδων και των αλγορίθμων που αναπτύχθηκαν στη διάρκεια αυτής της εργασίας, καθώς και των πειραματικών αποτελεσμάτων αυτών εμπεριέχεται πλέον στα [61, 62] και στα [30, 60].

1.4 Διάρθρωση της Εργασίας

Στο κεφάλαιο 2 θα παρουσιαστεί η βασική θεωρία του πλαισίου λήψης αποφάσεων υπό αβεβαιότητα σε μία κατάσταση, σε περιβάλλοντα για τα οποία οι στατιστικές των επιβραβεύσεων και κατ΄ επέκταση των αλληλεπιδράσεων παραμένουν αμετάβλητες. Αρχικά θα γίνει μαθηματική περιγραφή του προβλήματος, ενώ στη συνέχεια θα περιγραφούν οι βασικότεροι αλγόριθμοι αλλά και μέθοδοι οι οποίες θεωρούνται χρήσιμες για την κατανόηση του πλαισίου. Στο κεφάλαιο 3 θα γίνει επέκταση σε μη στατικά περιβάλλοντα, και θα παρουσιαστούν οι αλγόριθμοι με τους οποίους θα γίνει η μετέπειτα συγκριτική αξιολόγηση. Στο κεφάλαιο 4 παρουσιάζεται η ανάπτυξη του προτεινόμενου υβριδικού προσαρμοστικού αλγορίθμου MLB-KF, με χρήση βιολογικά εμπνευσμένης μετα-μάθησης (meta-learning for bandits) και φίλτρα Kalman (KF). Στο ίδιο κεφάλαιο παρουσιάζονται τα πειραματικά συγκριτικά αποτελέσματα μεταξύ των πιο σημαντικών εκ των προσαρμοστικών αλγόριθμων για λήψη αποφάσεων υπό

αβεβαιότητα σε μία κατάσταση. Στο κεφάλαιο 5 παρουσιάζεται το θεωρητικό υπόβαθρο του πλαισίου Μαρκοβιανών διαδικασιών λήψης αποφάσεων και ενισχυτικής μάθησης, καθώς και δύο σημαντικές προσεγγίσεις για στατικά περιβάλλοντα και χώρους παραμετροποιήσιμων διακριτών δράσεων. Στο κεφάλαιο 6 παρουσιάζεται η ανάπτυξη του νέου αλγορίθμου ενισχυτικής μάθησης για μη στατικά περιβάλλοντα, με χρήση εξειδικευμένης εξερεύνησης ανά κατάσταση. Στο ίδιο κεφάλαιο παρουσιάζονται τα πειραματικά αποτελέσματα του αλγορίθμου σε μία πληθώρα προβλημάτων. Στο κεφάλαιο 7 γίνεται τελική αξιολόγηση της εργασίας και συζήτηση για νέες κατευθύνσεις.

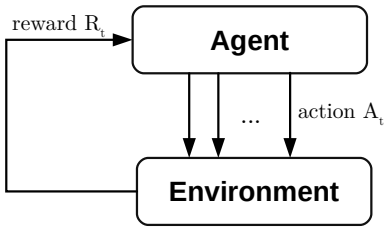
Κεφάλαιο 2

Μάθηση σε Στατικά Περιβάλλοντα μίας Κατάστασης

Το κύριο θέμα ενδιαφέροντος που θα αναλυθεί σε αυτό το κεφάλαιο είναι η στρατηγική λήψης αποφάσεων και δράσεων σε στατικά αλλά και σε δυναμικά περιβάλλοντα μίας κατάστασης. Βρισκόμενοι σε αυτή τη μία και μοναδική κατάσταση θεωρούμε ότι υπάρχει ένα πεπερασμένο σύνολο δράσεων επιλογής. Κάθε δράση δημιουργεί αλληλεπιδράσεις άγνωστης δυναμικής με το περιβάλλον και επιστέφεται μία στοχαστική παρατήρηση η οποία μπορεί να ποσοτικοποιηθεί ώστε να αντιστοιχεί σε μία τυχαία μεταβλητή κέρδους. Το ζητούμενο είναι η εύρεση της βέλτιστης πολιτικής λήψης αποφάσεων (στοχαστικής ή ντετερμινιστικής) η οποία μεγιστοποιεί τα αθροιστικά κέρδη.

Η βασική μαθηματική θεωρία που θα χρησιμοποιηθεί έχει τις ρίζες της από τον William R. Thompson το 1933 σε ένα δημοσιευμένο άρθρο της Biomertika [57], ο οποίος μελετούσε την αποτελεσματικότητα μίας προσαρμοστικής ιατρικής φαρμακευτικής αγωγής με βάση την τρέχουσα επίδραση του φαρμάκου, έναντι μίας προκαθορισμένης αγωγής η οποία ήταν η συνήθης την εποχή εκείνη. Στη συνέχεια οι Mosteller και Bush το 1950 μελέτησαν τους μηχανισμούς μάθησης σε βιολογικούς οργανισμούς, τοποθετώντας ένα εργαστηριακό ποντίκι στο κάτω άκρο ενός λαβύρινθου σχήματος-Τ, αλλάζοντας την τοποθεσία μίας τροφικής επιβράβευσης στα αριστερά ή δεξιά του λαβύρινθου και παρατηρώντας τις αποφάσεις που λάμβανε (στροφή αριστερά ή στροφή δεξιά) και τις σχέσεις αυτών με την ιστορία των δράσεων και των εκβάσεων.

Για τη μελέτη των αντίστοιχων μηχανισμών σε ανθρώπους, χρησιμοποιήθηκαν στη συνέχεια δύο ή περισσότερες μηχανές επιστροφής στοχαστικού τύπου χρηματικής ανταμοιβής (παρόμοιες με τους γνωστούς «κουλοχέρηδες», τις μηχανές τυχερών παιγνίων που υπάρχουν μέχρι και σήμερα στα καζίνο), όπου η κάθε μηχανή μπορούσε να επιστρέψει στον παίκτη ένα χρηματικό έπαθλο και η συνάρτηση πυκνότητας (ή μάζας) πιθανότητας κερδών από την οποία δειγματοληπτούσε ήταν άγνωστη στον παίκτη. Λόγω της ομοιότητας της πειραματικής διάταξης με αυτή των μηχανών τυχερών παιγνίων, τα προβλήματα είναι γνωστά στην παγκόσμια βιβλιογραφία με το όνομα multi-armed bandits (MAB), η ελληνική απόδοση των οποίων δεν είναι ακριβής και αν και θα επιλεχθεί ο όρος «στοχαστικές μηχανές επιβράβευσης πολλα-



Σχήμα 2.1: Στοχαστική μηχανή επιβράβευσης πολλαπλών επιλογών (MAB). Η μηχανή μάθησης επιλέγει μία δράση από ένα διακριτό σύνολο και το περιβάλλον επιστρέφει μία επιβράβευση.

πλών επιλογών», ή απλούστερα «μηχανές επιβράβευσης», θα χρησιμοποιούμε πολλές φορές το ακρωνύμιο MAB ή την αγγλική απόδοση.

2.1 Multi-Armed Bandits (MAB)

Η λήψη αποφάσεων υπό αβεβαιότητα είναι ένας από τους βασικούς τομείς εφαρμογής της θεωρίας ανάλυσης των MABs. Μεγάλες τεχνολογικές εταιρείες χρησιμοποιούν αλγορίθμους MAB για εφαρμογές ρύθμισης της διεπαφής ιστού, όπως για την πρόταση ειδήσεων ενδιαφέροντος για τον κάθε χρήστη ξεχωριστά, την εμφάνιση διαφημιστικού υλικού με σκοπό την μεγιστοποίηση της πιθανότητας ο χρήστης να δείξει ενδιαφέρον για τις προτάσεις αυτές, ή ακόμα και ως δομικό στοιχείο πιο σύνθετων μεθόδων όπως για παράδειγμα στη χρήση τους για την Monte-Carlo αναζήτηση σε δέντρα (MCTS) που έπαιξε πολύ σημαντικό ρόλο στις πρόσφατες επιτυχίες του υπολογιστή AlphaGo [33, 52]. Επιπρόσθετα, ο μαθηματικός φορμαλισμός των προβλημάτων MAB έχει μία αρκετά κομψή αλλά και πλούσια δομή η οποία μπορεί να γεφυρώσει διαφορετικούς τομείς των μαθηματικών. Ίσως αυτοί είναι και μερικοί από τους λόγους που το πλήθος των δημοσιεύσεων των επιστημονικών άρθρων που αφορούν αλγορίθμους ή εφαρμογές MAB ανά έτος, δείχνει να αυξάνεται γραμμικά την τελευταία δεκαετία.

Ας υποθέσουμε ότι έχουμε στη διάθεσή μας μία τέτοια στοχαστική μηχανή επιβράβευσης δύο επιλογών, ή αλλιώς ένα 2-armed bandit με δυνατότητα επιλογής του αριστερού ή του δεξιού «άκρου» και γνωρίζουμε τις 5 παρακάτω τελευταίες εκβάσεις επιστροφής κέρδους μετά από επιλογή του κάθε άκρου:

Αριστερό άκρο	:	10, 0, 0, 10, 10
Δεξί άκρο	:	15, 0, 0, 0, 0

Με βάση αυτή και μόνο την εμπειρία φαίνεται ότι το αριστερό άκρο αποδίδει περισσότερο καθώς το μέσο κέρδος του είναι €6 ανα γύρο, ενώ το μέσο κέρδος του δεξιού άκρου είναι €3. Τα ενδιαφέροντα ερωτήματα που τίθενται ωστόσο είναι αρκετά. Στην περίπτωση που έχουμε στη διάθεσή μας 20 επιλογές, πώς θα επιλέγαμε τα άκρα στη συνέχεια; Θα συνεχίζαμε να επιλέγουμε το αριστερό άκρο χωρίς να λαμβάνουμε υπόψιν το δεξί, ή μήπως θα θα επιλέγαμε

το δεξιό άκρο μερικές φορές για να αποκτήσουμε καλύτερη στατιστική σημαντικότητα ως προς τα αναμενόμενα κέρδη του, παρόλο που δεν φαίνεται να είναι η καλύτερη επιλογή; Στην πρώτη περίπτωση οι δράσεις μας θα ήταν αξιοποιητικές (exploitative) ενώ στην δεύτερη εξερευνητικές (explorative).

Είναι φανερό λοιπόν ότι η λήψη αποφάσεων υπό αβεβαιότητα με ανατροφοδότηση των εκβάσεων σε πραγματικό χρόνο συμπεριλαμβάνει το δίλημμα αξιοποίησης της υπάρχουσας γνώσης, ή της εξερεύνησης για συλλογή περισσότερης γνώσης. Η εύρεση μίας δυναμικά εξελισσόμενης αναλογίας μεταξύ εξερευνητικών και αξιοποιητικών δράσεων βρίσκεται στην καρδιά όλων των MAB προβλημάτων.

2.1.1 Μαθηματικός Φορμαλισμός των προβλημάτων MAB

Ένα πρόβλημα MAB μπορεί να θεωρηθεί ως μία ακολουθία παιγνίων μεταξύ του *μαθητευόμενου* (learner) και του *περιβάλλοντος* (environment). Ο μαθητευόμενος, ο οποίος στα τεχνητά προβλήματα που θα μας απασχολήσουν είναι μία μηχανή μάθησης, αλληλεπιδρά με το περιβάλλον για T γύρους, όπου $T \in \mathbb{N}^+$ είναι ο *ορίζοντας* του προβλήματος. Σε κάθε γύρο η μηχανή επιλέγει κάποια δράση $A_t \in \mathcal{A}$ από ένα σύνολο δράσεων $\mathcal{A}$ (προς το παρόν διακριτό και πεπερασμένο) και το περιβάλλον επιστρέφει μία ανταμοιβή $R_t \in \mathbb{R}$. Η επιλογή της δράσης A_t κάθε χρονική στιγμή t μπορεί να εξαρτάται μόνο από την *ιστορία* $\mathcal{H}$ (history), η οποία ορίζεται ως το σύνολο των δράσεων και των εκβάσεων έως εκείνη τη χρονική στιγμή, με $H_{t-1} = \{A_1, R_1, \ldots, A_{t-1}, R_{t-1}\}$. Πιο συγκεκριμένα, η δράση A_t επιλέγεται μέσω της *πολιτικής* (policy) $\pi \in \Pi$, η οποία μπορεί να περιγραφεί ως μία απεικόνιση από το χώρο ιστοριών στο χώρο δράσεων $\pi : \mathcal{H} \mapsto \mathcal{A}$, αν και στη γενικότερη περίπτωση θα έχει το ρόλο της συνάρτησης μάζας πιθανότητας από την οποία δειγματοληπτούνται οι λήψεις αποφάσεων με $\pi(a) = \mathbb{P}[A_t = a]$ για $a \in \mathcal{A}$. Ο στόχος της μηχανής μάθησης θα είναι η εύρεση της βέλτιστης πολιτικής $\pi^\star$ η οποία μεγιστοποιεί τις αθροιστικές ανταμοιβές από την αρχή του παιγνίου έως το τέλος του ορίζοντα, με

$$\pi^\star = \underset{\pi}{\operatorname{argmax}} \sum_{t=1}^{T} \mathbb{E}[R_t]$$

ενώ η μετρική που χρησιμοποιείται κυρίως για την αξιολόγηση της διαδικασίας ονομάζεται *μεταμέλεια* (regret), ο ορισμός της οποίας ακολουθεί στις επόμενες γραμμές.

Ορισμός 2.1. *Η μεταμέλεια (regret) μίας ακολουθίας δράσεων σε σχέση με μία πολιτική αναφοράς π, είναι η διαφορά μεταξύ της συνολικά εκτιμώμενης επιβράβευσης που προκύπτει σε T γύρους ακολουθώντας την πολιτική π και της συνολικής εκτιμώμενης επιβράβευσης από τις δεδομένες δράσεις. Η μεταμέλεια των δράσεων σε σχέση με ένα σύνολο πολιτικών Π ορίζεται ως η μέγιστη μεταμέλεια σε σχέση με κάθε πολιτική $\pi \in \Pi$.*

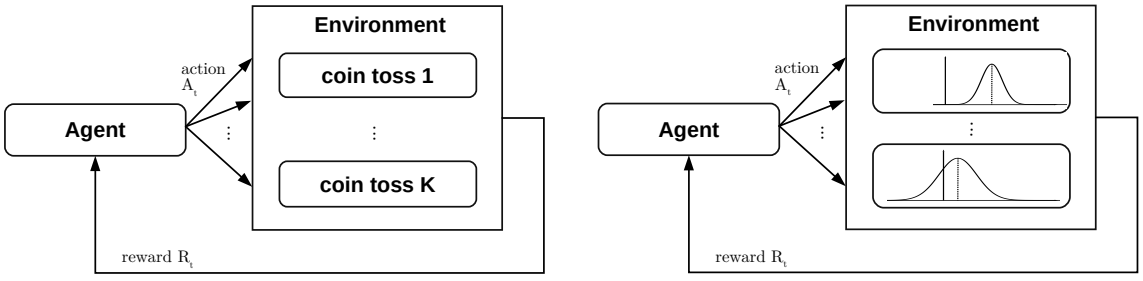
Μία σημαντική κατηγορία που θα μας απασχολήσει είναι τα *στοχαστικά Bernoulli bandits*, ενώ στη συνέχεια θα θεωρούμε ότι το σύνολο πολιτικών Π είναι αρκετά μεγάλο ώστε να συμπεριλαμβάνει την βέλτιστη πολιτική $\pi^\star$. Σε αυτά τα περιβάλλοντα θεωρούμε το σύνολο δράσεων $\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$ ενώ οι επιβραβεύσεις $R^a \in \{0, 1\}$ που επιστρέφει το περιβάλλον έπειτα από κάθε δράση $a \in \mathcal{A}$ είναι δυαδικές και ακολουθούν την κατανομή Bernoulli $R^a \sim \mathcal{B}(1, \mu_a)$, έτσι ώστε $\mu_a \in [0, 1]$ με $\mu_a = \mathbb{E}[R^a]$. Γενικότερα, μπορούμε να θεωρήσουμε ότι το περιβάλλον έπειτα από κάθε αλληλεπίδραση επιστρέφει την επιβράβευση R, οπότε $\mu_a = \mathbb{E}[R|a]$.

Το στοχαστικό bandit τύπου Bernoulli μπορεί να περιγραφεί πλήρως από το διάνυσμα $\boldsymbol{\mu} \in [0, 1]^K$ εντός του μοναδιαίου K-διάστατου υπερκύβου, όπου $\boldsymbol{\mu} = [\mu_{a_1}, \mu_{a_2}, \ldots, \mu_{a_K}]^T$, και συνεπώς η βέλτιστη απόφαση δράσης $a^\star$ θα είναι στατική και ίση με $a^\star = \text{argmax}_{a \in \mathcal{A}} \mu_a$. Το διάνυσμα $\boldsymbol{\mu}$ δεν είναι γνωστό στην μηχανή μάθησης, αλλά εάν υποθέσουμε ότι είναι γνωστό στον αξιολογητή της διαδικασίας (συνήθως ο αξιολογητής είμαστε εμείς) η συνολική μεταμέλεια L_T για ορίζοντα T θα είναι ίση με

$$L_T = T \max_{a \in \mathcal{A}} \mu_a - \mathbb{E}\left[\sum_{t=1}^{T} R_t\right] \tag{2.1}$$

Είναι προφανές ότι όσο μικρότερη είναι η μεταμέλεια τόσο πιο αποδοτική είναι μία μηχανή μάθησης, ενώ ένα από βασικά ερωτήματα είναι η εξάρτησή της από το μέγεθος του ορίζοντα. Για έναν σχετικά καλό αλγόριθμο μάθησης θα ισχύει ότι $L_T = O(T)$, παρόλα αυτά στα στοχαστικά και στατικά περιβάλλοντα (stochastic stationary bandits) που μας αφορούν (σε αυτό το κεφάλαιο) μπορεί να επιτευχθεί και καλύτερο φράγμα ώστε $L_T = O(logT)$.

Στην οικογένεια των στοχαστικών και στατικών MABs οι κατανομές πυκνότητας πιθα-



Σχήμα 2.2: Αριστερά: Διάγραμμα μηχανών επιβράβευσης τύπου Bernoulli. Η μηχανή μάθησης επιλέγει μία δράση από το διαθέσιμο σύνολο και το περιβάλλον επιστρέφει 1 ή 0 εκτελώντας ρίψη ενός μεροληπτικού νομίσματος. Δεξιά: Διάγραμμα μηχανών επιβράβευσης Gaussian τύπου. Το περιβάλλον μετά από κάθε δράση δειγματοληπτεί την επιβράβευση από την αντίστοιχη Gaussian κατανομή.

νότητας ανταμοιβών καθορίζονται μόνο από την εκάστοτε δράση (έχουμε μία χρονοσταθερή κατανομή για κάθε δράση) και είναι ανεξάρτητες από την ιστορία $\mathcal{H}$ των δράσεων-εκβάσεων αλλά και μεταξύ τους. Οι κατανομές αυτές μπορούν επίσης να έχουν οποιαδήποτε παραμετρική ή μη παραμετρική μορφή. Για παράδειγμα οι ανταμοιβές R^a της δράσης a μπορούν να ακολουθούν Gaussian κατανομή με $R^a \sim \mathcal{N}(\mu_a, \sigma_a^2)$ αντί για Bernoulli κατανομή $\mathcal{B}(1, \mu_a)$ που περιγράφηκε νωρίτερα (Σχήμα 2.2). Οι τρεις βασικές διαφορετικές οικογένειες που υπάρχουν, προκύπτουν από τις διαφορετικές υποθέσεις που μπορεί να κάνει κανείς για τη φύση της γεννήτριας επιβραβεύσεων και είναι οι ακόλουθες:

Στοχαστικά MABs: Σε αυτή την οικογένεια προβλημάτων, κάθε δράση $a \in \mathcal{A}$ σχετίζεται με μια άγνωστη συνάρτηση πυκνότητας πιθανότητας $\mathcal{R}^a$, ενώ η αντίστοιχη επιβράβευση $R^a \sim \mathcal{R}^a$ δειγματοληπτείται ανεξάρτητα και όμοια κατανεμημένα (i.i.d) από αυτήν. Οι αλγόριθμοι που λειτουργούν βέλτιστα σε αυτά τα προβλήματα ονομάζονται UCB (Upper Confidence Bound) [3] και θα περιγραφούν στη συνέχεια.

Ανταγωνιστικά MABs: Σε αυτή την οικογένεια δεν υπάρχουν υποθέσεις και περιορισμοί για την υποκείμενη γεννήτρια των επιβραβεύσεων. Ως παράδειγμα οι ανταμοιβές μπορεί να εξαρτώνται από την ιστορία των δράσεων/εκβάσεων $\mathcal{H}$ ή να ελέγχονται από κάποιον «ανταγωνιστή». Η κατηγορία αλγορίθμων που επιτυγχάνει την καλύτερη επίδοση είναι οι αλγόριθμοι τύπου Exp3 [46].

Μαρκοβιανά MABs: Σε αυτή την οικογένεια οι επιβραβεύσεις που προκύπτουν μετά από κάθε δράση ακολουθούν κάποια Μαρκοβιανή διαδικασία ενός άγνωστου υποκείμενου χώρου κατάστασης. Τα εργαλεία και οι αλγόριθμοι που χρησιμοποιούνται σε αυτή την οικογένεια προβλημάτων είναι διαφορετικού τύπου, με βέλτιστη στρατηγική τη χρήση των δεικτών του Gittins [20].

Σε αυτή την εργασία θα μας απασχολήσει η πρώτη κατηγορία, ενώ στο επόμενο κεφάλαιο θα παρουσιαστούν οι βασικότερες μέθοδοι που προσεγγίζουν την περίπτωση χρονομεταβλητών κατανομών.

Γενικεύοντας την έννοια της μεταμέλειας της εξίσωσης (2.1), η οποία αφορά μόνο την περίπτωση των στοχαστικών MABs, θεωρούμε ότι ο αλγόριθμος μάθησης μπορεί κάθε χρονική στιγμή t να επιλέξει μία δράση A_t με στοχαστικό ή ντετερμινιστικό τρόπο ενώ το περιβάλλον επιστρέφει την ανταμοιβή $R_{A_t,t}$. Η μεταμέλεια για ορίζοντα T θα είναι

$$L_T = \max_{a \in \mathcal{A}} \sum_{t=1}^{T} R_{a,t} - \sum_{t=1}^{T} R_{A_t,t}$$

Όπως ήδη αναφέραμε, στη γενική περίπτωση οι επιβραβεύσεις $R_{A_t,t}$ του περιβάλλοντος

αλλά και οι δράσεις A_t της μηχανής μάθησης μπορεί να είναι στοχαστικές, κάτι που οδηγεί σε δύο διαφορετικούς ορισμούς για την μέση μεταμέλεια. Ο πρώτος αφορά *εκτιμώμενη μεταμέλεια (expected regret)*

$$\mathbb{E}[L_T] = \mathbb{E}\left[\max_{a\in\mathcal{A}} \sum_{t=1}^{T} R_{a,t} - \sum_{t=1}^{T} R_{A_t,t}\right] \tag{2.2}$$

ενώ ο δεύτερος πιο χαλαρός ορισμός αφορά την *ψευδο-μεταμέλεια (pseudo-regret)* και είναι αυτός που οδηγεί στην εξίσωση (2.1) στην περίπτωση των στοχαστικών MABs, με

$$\bar{L}_T = \max_{a\in\mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} R_{a,t} - \sum_{t=1}^{T} R_{A_t,t}\right] \tag{2.3}$$

Ορίζοντας το χάσμα δράσης Δ_a κάθε δράσης a ως $\Delta_a = \mu^\star - \mu_a$, με $\mu^\star = \max_a\{\mu_a\}$ και $N_{a,t}$, το πλήθος των δράσεων a από την αρχή του χρόνου έως την χρονική στιγμή t

$$N_{a,t} = \sum_{s=1}^{t} \mathbb{I}\{A_s = a\}$$

όπου $\mathbb{I}\{\cdot\}$ είναι η συνάρτηση δεικτοδότησης η οποία επιστρέφει μονάδα εαν το όρισμα είναι αληθές και μηδέν αν το όρισμα είναι ψευδές, τότε η μεταμέλεια L_T μπορεί να ορισθεί με το παρακάτω Λήμμα διάσπασης [3],

Λήμμα 2.1. *(Διάσπαση Μεταμέλειας) Για κάθε πολιτική π που ακολουθείται σε μία στοχαστική μηχανή επιβράβευσης πολλαπλών επιλογών (MAB) με ορίζοντα T, η μεταμέλεια (regret) L_T της πολιτικής π ικανοποιεί τη σχέση*

$$L_T = \sum_{a\in\mathcal{A}} \Delta_a \mathbb{E}[N_{a,T}] \tag{2.4}$$

Το παραπάνω λήμμα φανερώνει ότι για να ελαχιστοποιηθεί η μεταμέλεια (ή ισοδύναμα να μεγιστοποιηθούν οι αθροιστικές ανταμοιβές) μίας μηχανής μάθησης, θα πρέπει να ελαχιστοποιηθεί το σταθμισμένο άθροισμα των δράσεων, όπου τα βάρη αντιστοιχούν στα χάσματα δράσης.

Απόδειξη του Λήμματος 2.1: Έστω ότι ο τελεστής $\mathbb{E}_t[\cdot]$ επιστρέφει την εκτίμηση υπό συνθήκη, με δεδομένη την ιστορία H_{t-1} των δράσεων/εκβάσεων έως τη χρονική στιγμή $t-1$ και την τελευταία εκτελούμενη δράση A_t, δηλαδή $E_t[\cdot] = E[\cdot|H_{t-1}, A_t]$, τότε

$$L_T = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} R_t\right] = T\mu^\star - \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}_t\left[R_t\right]\right] = T\mu^\star - \sum_{t=1}^{T} \mathbb{E}\left[\mu_{A_t}\right]$$

$$= \sum_{t=1}^{T} \mathbb{E}\left[\Delta_{A_t}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{A_t}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}} \mathbb{I}\{A_t = a\}\Delta_a\right]$$

$$= \left[\sum_{a\in\mathcal{A}} \Delta_a \sum_{t=1}^{T} \mathbb{I}\{A_t = a\}\right] = \mathbb{E}\left[\sum_{a\in\mathcal{A}} \Delta_a N_{a,T}\right] = \sum_{a\in\mathcal{A}} \Delta_a \mathbb{E}\left[N_{a,T}\right]$$

Για την ελαχιστοποίηση της ψευδο-μεταμέλειας, το ζητούμενο είναι να βρεθεί η δράση a με την υψηλότερη μέση επιβράβευση $\mu^\star$. Γι' αυτόν το σκοπό υπάρχουν αρκετές στρατηγικές μάθησης, οι περισσότερες των οποίων χρησιμοποιούν τους δειγματικούς μέσους των επιβραβεύσεων ώστε να εκτιμήσουν τους πραγματικούς μέσους μ_a. Ένα τέτοιο παράδειγμα είναι ο άπληστος (greedy) αλγόριθμος, ο ϵ-άπληστος (ϵ-greedy) και ο αλγόριθμος επιλογής δράσεων με soft-max Boltzmann κατανομή (αναφέρεται και ως Gibbs).

2.1.2 Άπληστοι και Στοχαστικοί Αλγόριθμοι

Η πρώτη και βασική ιδέα είναι η εκτίμηση $\hat{\mu}_{a,t}$ των πραγματικών μέσων επιβραβεύσεων μ_a των δράσεων κάθε χρονική στιγμή t, έτσι ώστε $\hat{\mu}_{a,t} \approx \mu_a$. Υποθέτοντας ότι η μηχανή μάθησης επιλέγει μία δράση A_t κάθε χρονική στιγμή t, ενώ το περιβάλλον επιστρέφει την επιβράβευση r_t, τότε η εκτίμηση μπορεί να γίνει με την απλή Monte-Carlo μέθοδο ως

$$\hat{\mu}_{a,t} = \frac{1}{N_{a,t}} \sum_{s=1}^{t} r_s \mathbb{I}\{A_s = a\}$$

όπου ο απλός greedy αλγόριθμος σε κάθε βήμα t επιλέγει την βέλτιστη δράση $a_t^\star$ για την οποία

$$a_t^\star = \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, \hat{\mu}_{a,t}$$

ωστόσο είναι εμφανές ότι λόγω της στοχαστικότητας του περιβάλλοντος ο greedy αλγόριθμος μπορεί να συνεχίσει να επιλέγει μια υποβέλτιστη δράση καθώς δεν εκτελεί εξερευνητικές δράσεις. Γι΄ αυτό το λόγο επιτυγχάνει γραμμική μεταμέλεια $L_T = O(T)$.

Μία καλύτερη προσέγγιση είναι αυτή του ϵ-greedy αλγορίθμου, όπου η μηχανή μάθησης επιλέγει τη βέλτιστη εκτιμώμενη δράση $a_t^\star$ με πιθανότητα $1-\epsilon$, ενώ επιλέγει μία τυχαία δράση με πιθανότητα ϵ, όπου $0 \leq \epsilon \leq 1$. Εάν ℓ_t είναι η στιγμιαία μεταμέλεια, έτσι ώστε $L_T = \sum_{t=1}^{T} \ell_t$, είναι εύκολο να παρατηρήσουμε ότι

$$\ell_t = \sum_{a \in \mathcal{A}} \mathbb{P}[A_t = a]\Delta_a = \frac{\epsilon}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \Delta_a \Rightarrow L_T = T\frac{\epsilon}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \Delta_a$$

δηλαδή επιτυγχάνει και αυτός γραμμική μεταμέλεια (αν και εμπειρικά αποδίδει καλύτερα από τον απλό greedy αλγόριθμο).

Μία ακόμα τροποποίηση που μπορεί να οδηγήσει σε καλύτερη επίδοση και συνήθως προτιμάται, είναι η *αισιόδοξη αρχικοποίηση* κατά την οποία η αρχικοποίηση των εκτιμήσεων $\hat{\mu}_a$ γίνεται σε κάποια υψηλή τιμή κοντά στο άνω όριο των επιβραβεύσεων (ή ίση). Η ανανέωση των εκτιμήσεων μπορεί να γίνεται σε κάθε βήμα, ώστε

$$\hat{\mu}_{A_t,t} = \hat{\mu}_{A_t,t-1} + \frac{1}{N_{A_t,t}}(r_t - \hat{\mu}_{A_t,t-1})$$

Η παραπάνω τεχνική ενθαρρύνει τις εξερευνητικές δράσεις κατά τα πρώτα στάδια μάθησης, κάτι το οποίο είναι χρήσιμο για τη συλλογή απαραίτητων δειγμάτων ώστε να αποφευχθεί ο εγκλωβισμός στη συνεχή λήψη κάποιας υποβέλτιστης επιλογής. Παρ΄ όλα αυτά το πρόβλημα δεν επιλύεται πλήρως, ούτε και αλλάζει η γραμμική εξάρτηση της μεταμέλειας από το μέγεθος του ορίζοντα (εμπειρικά όμως αποδίδει καλύτερα).

Εναλλακτικά, οι αποφάσεις μπορούν να λαμβάνονται βάσει της συνάρτησης μάζας πιθανότητας που προκύπτει με χρήση της soft-max Boltzmann συνάρτησης, έτσι ώστε

$$\mathbb{P}[A_t = a] = \frac{\exp\{\beta\hat{\mu}_{a,t}\}}{\sum_{a' \in \mathcal{A}} \exp\{\beta\hat{\mu}_{a',t}\}} \tag{2.5}$$

όπου β είναι μία παράμετρος αντίστροφης θερμοκρασίας εμπνευσμένη από την τεχνική προσομοιωμένης ανόπτησης. Η λογική είναι ότι για μικρές τιμές του β (ή αντίστοιχα μεγάλες τιμές θερμοκρασίας) η τυχαιότητα επιλογής δράσης είναι μεγάλη, ενώ για μεγάλες τιμές ο αλγόριθμος τείνει να γίνει ντετερμινιστικός. Συγκεκριμένα για $\beta = 0$ η πιθανότητα επιλογής

δράσης είναι ομοιόμορφη ενώ για $\beta \to \infty$ ο αλγόριθμος καταλήγει στον greedy. Η ουσιαστική διαφορά με τον ϵ-greedy είναι ότι ο τελευταίος δεν λαμβάνει υπόψιν την εκτιμώμενη αξία κάθε δράσης κατά την εξερεύνηση σε αντίθεση με τη soft-max μεθόδο η οποία εμπεριέχει μία μόνιμη αλλά δυναμικά εξελισσόμενη ιεραρχική εξερεύνηση λαμβάνοντας υπόψιν την απόδοση κάθε δράσης γι΄ αυτόν τον σκοπό.

Μια ακόμα καλύτερη βελτίωση είναι η σταδιακή μείωση των εξερευνητικών δράσεων. Αυτό επιτυγχάνεται επιλέγοντας μία φθίνουσα ακολουθία $\{\epsilon_1, \epsilon_2, \dots \epsilon_n\}$ για την πιθανότητα ϵ_t να πραγματοποιηθεί μία εξερευνητική δράση. Συγκεκριμένα, ο αλγόριθμος decaying ϵ-greedy χρησιμοποιεί το παρακάτω πλάνο εξερεύνησης

$$\epsilon_t = \min\left\{1, \frac{c|\mathcal{A}|}{d^2 t}\right\}, \text{ όπου } c > 0 \text{ και } d = \min_{a \in \mathcal{A}\setminus a^\star} \Delta_a$$

με τον περιορισμό ότι το χάσμα δράσης της πρώτης υποβέλτιστης δράσης θα πρέπει να είναι γνωστό και στην μηχανή μάθησης (χωρίς ωστόσο να χρειάζεται να γνωρίζει ποια δράση αφορά). Ο αλγόριθμος αυτός επιτυγχάνει ασυμπτωτικά λογαριθμική μεταμέλεια $L_T = O(logT)$. Με μία πιο λεπτομερή διερεύνηση είναι εμφανές ότι η επίδοση εξαρτάται από την ομοιότητα των κατανομών πιθανότητας επιβράβευσης μεταξύ της βέλτιστης δράσης $\mathcal{R}^{a^\star}$ και των υπολοίπων κατανομών $\mathcal{R}^a$ των υποβέλτιστων δράσεων $a \in \mathcal{A}\setminus a^\star$. Έτσι, η δυσκολία ενός προβλήματος καθορίζεται από τα χάσματα Δ_a αλλά και από τις ομοιότητες των κατανομών επιβράβευσης. Χρησιμοποιώντας την απόκλιση Kullback-Leibler οδηγούμαστε στο παρακάτω θεώρημα [37]

Θεώρημα 2.1. *(Lai και Robbins) Η ασυμπτωτική μεταμέλεια μίας μηχανής μάθησης είναι τουλάχιστον λογαριθμική ως προς τον αριθμό των βημάτων*

$$\lim_{T\to\infty} L_T = \log T \sum_{a \in \mathcal{A}\setminus a^\star} \frac{\Delta_a}{KL(\mathcal{R}^a || \mathcal{R}^{a^\star})}$$

συνεπώς ο αλγόριθμος decaying ϵ_t-greedy επιτυγχάνει βέλτιστη επίδοση ασυμπτωτικά, ενώ το ζητούμενο είναι να επιτευχθεί η ίδια επίδοση χωρίς να είναι γνωστά τα χάσματα δράσεων.

2.1.3 Αλγόριθμοι Αισιοδοξίας στο Μέτωπο της Αβεβαιότητας (UCB)

Οι αλγόριθμοι που αναφέρθηκαν παραπάνω χρησιμοποιούν αποκλειστικά τους δειγματικούς μέσους των επιβραβεύσεων κάθε δράσης για τη στρατηγική λήψης απόφασης. Η οικογένεια αλγόριθμων UCB (Upper Confidence Bound) [3] χρησιμοποιεί επιπλέον των δειγματικών μέσων $\hat{\mu}_{a,t}$, τα διαστήματα εμπιστοσύνης αυτών, ενώ σε κάθε βήμα t επιλέγεται η δράση για την οποία παρατηρείται το μεγαλύτερο άνω όριο εκ των διαστημάτων.

Το άνω όριο του διαστήματος εμπιστοσύνης για τη μέση επιβράβευση κάποιας δράσης μπορεί να είναι υψηλό για δύο βασικούς λόγους. Κατά πρώτον στην περίπτωση που η δράση

αυτή δεν έχει επιλεχθεί αρκετές φορές, θα υπάρχει μεγάλη αβεβαιότητα για την έκβαση, και το άνω όριο θα είναι υψηλό (ακόμα και στην περίπτωση που ο δειγματικός μέσος των επιβραβεύσεων έχει χαμηλή τιμή). Κατά δεύτερον στις περιπτώσεις που ο ίδιος ο δειγματικός μέσος των επιβραβεύσεων έχει υψηλή τιμή τότε και το άνω όριο του διαστήματος εμπιστοσύνης θα είναι υψηλό. Στην πρώτη περίπτωση ο αλγόριθμος θα εκτελούσε εξερευνητικές δράσεις, ενώ στη δεύτερη περίπτωση αξιοποιητικές. Η λογική λοιπόν της *αισιοδοξίας στο μέτωπο της αβεβαιότητας* βασίζεται στο ότι δράσεις οι οποίες, από την υπάρχουσα πληροφορία, δεν φαίνεται να είναι βέλτιστες, αλλά για τις οποίες είμαστε αβέβαιοι σε βαθμό που η πιθανότητα να είναι βέλτιστες δεν είναι αμελητέα, πρέπει να επιλεχθούν.

Το ζητούμενο είναι σε κάθε βήμα t να υπολογίζεται για κάθε δράση a ένα εύρος $u_{a,t}$ τέτοιο ώστε η πιθανότητα $\mathbb{P}[\mu_a \leq \hat{\mu}_{a,t} + u_{a,t}]$ να είναι ικανοποιητικά μεγάλη. Η επιλογή δράσης σε κάθε βήμα γίνεται ντετερμινιστικά με

$$A_t = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \{\hat{\mu}_{a,t} + u_{a,t}\}$$

όπου για τον υπολογισμό του $u_{a,t}$ θα γίνει χρήση της ανισότητας του Hoeffding (επίσης αναφέρεται και ως φράγμα Chernoff) όπως αναφέρεται στο παρακάτω θεώρημα [24].

Θεώρημα 2.2. *(Ανισότητα του Hoeffding - φράγμα Chernoff) Εάν $X_1, \ldots, X_t$ είναι ανεξάρτητες και όμοια κατανεμημένες μεταβλητές (i.i.d) στο διάστημα $[0, 1]$, και $\bar{X}_t = \frac{1}{t} \sum_{\tau=1}^{t} X_\tau$ είναι ο δειγματικός τους μέσος, τότε*

$$\mathbb{P}[\mathbb{E}[X] > \bar{X}_t + u] \leq \exp\{-2tu^2\}$$

Η παραπάνω ανισότητα είναι αρκετά χρήσιμη καθώς γνωρίζοντας για κάθε δράση a το πόσες φορές $N_{a,t}$ έχει επιλεγεί, αλλά και τον δειγματικό μέσο $\hat{\mu}_{a,t}$, προκύπτει ότι

$$\mathbb{P}[\mu_a > \hat{\mu}_{a,t} + u_{a,t}] \leq \exp\{-2N_{a,t}u_{a,t}^2\}$$

συνεπώς επιλέγοντας κάποια τιμή πιθανότητας p αρκετά μικρή ώστε $\mathbb{P}[\mu_a > \hat{\mu}_{a,t} + u_{a,t}] = p$ και χρησιμοποιώντας τη σχέση του Hoeffding μπορούμε να βρούμε το επιθυμητό εύρος $u_{a,t}$. Εάν μάλιστα η τιμή της πιθανότητας p φθίνει ως συνάρτηση του χρόνου ακολουθώντας τη σχέση σχέση $p(t) = t^{-4}$ ο αλγόριθμος καταλήγει στον $UCB1$ και η λήψη αποφάσεων γίνεται με βάση τον παρακάτω ντετερμινιστικό τρόπο

$$A_t = \text{argmax}\left\{\hat{\mu}_{a,t} + \sqrt{\frac{2\log t}{N_{a,t}}}\right\}$$

ο οποίος αποδεικνύεται [3] ότι επιτυγχάνει λογαριθμική μεταμέλεια, παρόμοια με τον decaying ϵ_t-greedy, χωρίς ωστόσο να χρησιμοποιεί τα χάσματα δράσεων.

Θεώρημα 2.3. *Ο αλγόριθμος Upper Confidence Bound επιτυγχάνει μεταμέλεια με ασυμπτωτικά λογαριθμική συμπεριφορά ως προς το μέγεθος του ορίζοντα*

$$\lim_{T\to\infty} L_T \leq 8\log T \sum_{a\in\mathcal{A}\setminus a^\star} \Delta_a$$

Κάποιες παραλλαγές του αλγορίθμου UCB1, όπως ο UCB2 και ο UCB-Tuned μπορεί να οδηγήσουν σε λίγο καλύτερα αποτελέσματα. Συγκεκριμένα στον UCB-Tuned ο υπολογισμός του εύρους $u_{a,t}$ γίνεται με τη σχέση

$$\tilde{u}_{a,t} = \sqrt{\frac{\log t}{N_{a,t}} \min\left\{\frac{1}{4}, \hat{V}_{a,t}\right\}}$$

όπου $\hat{V}_{a,t}$ είναι μία εκτίμηση για το άνω όριο της διασποράς των επιβραβεύσεων της δράσης a, η οποία μπορεί να υπολογιστεί προσθέτοντας την διασπορά των δειγμάτων στην προηγούμενη σχέση για το εύρος $u_{a,t}$ που χρησιμοποιεί ο UCB1, ως

$$\hat{V}_{a,t} = \frac{1}{N_{a,t}} \sum_{s=1}^{t} r_s^2 \mathbb{I}\{A_s = a\} - \hat{\mu}_{a,t}^2 + \sqrt{\frac{2\log t}{N_{a,t}}}$$

Ο UCB-Tuned επιτυγχάνει καλύτερα εμπειρικά αποτελέσματα από τους UCB1 και UCB2 όσον αφορά τη συχνότητα με την οποία επιλέγει τη βέλτιστη δράση. Επιπρόσθετα, όπως αναφέρεται στο [3], παρατηρείται ότι είναι πιο εύρωστος στις περιπτώσεις κατανομών επιβραβεύσεων με μεγάλη διασπορά.

2.1.4 Δείκτες του Gittins

Ως *πολιτικές δείκτη* χαρακτηρίζονται οι στρατηγικές κατά τις οποίες υπολογίζεται ένας δείκτης για κάθε δράση και η απόφαση λαμβάνεται με βάση τον μέγιστο δείκτη χωρίς παράλληλα να απαιτείται γνώση της πληροφορίας για τις τιμές των δεικτών από δράση σε δράση. Οι δείκτες του Gittins [20] είναι μία τέτοια πολιτική όπου το ζητούμενο είναι η ελαχιστοποίηση

της τροποποιημένης συνολικής μεταμέλειας όπως ορίζεται παρακάτω,

$$\tilde{L}_T = \sum_{t=1}^{T} \gamma^{T-t}(\mu_{\max} - \mu_{A_t}) \tag{2.6}$$

όπου $0 < \gamma < 1$ μία παράμετρος επιλογής έτσι ώστε η μεταμέλεια των δράσεων να έχει εκθετική μείωση σημαντικότητας αντίστοιχη με το πόσο «μακριά» στο παρελθόν πραγματοποιήθηκε η κάθε δράση. Έχοντας γνώση (ή εκτίμηση) της πρότερης πιθανότητας για τις μέσες αξίες των δράσεων $P(\mu_a)$, οι δείκτες Gittins $\kappa(P(\mu_a))$ κάθε δράσης υπολογίζονται μέσω Bayesian λογικής και είναι

$$\kappa(P(\mu_a)) = \sup_{\tau} \left[\frac{\int \mathbb{E}_{\mu_a}[R_{a,1}] + \sum_{t=1}^{\tau-1} \gamma^t \mathbb{E}[R_{a,t}|R_{a,1}, \dots R_{a,t-1}] dP(\mu_a)}{\int \sum_{t=0}^{\tau} \gamma^t dP(\mu_a)} \right]$$

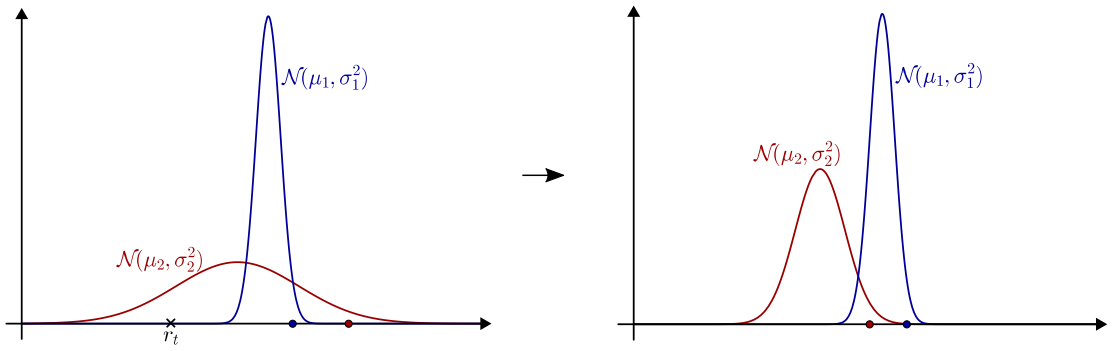
παραπέμποντας στο [20] για την προέλευση της παραπάνω σχέσης. Η εύρεσή τους (η οποία μπορεί να γίνει χρησιμοποιώντας δυναμικό προγραμματισμό) επιλύει πλήρως το πρόβλημα του διλήμματος εξερεύνησης-αξιοποίησης καθώς ελαχιστοποιεί τη σταθμισμένη μεταμέλεια $\tilde{L}_T$. Για $\gamma \to 1$ οι δείκτες του Gittins αναμένεται ότι επιλύουν το κλασσικό στοχαστικό MAB πρόβλημα. Παρ' όλα αυτά δυστυχώς τα πράγματα είναι διαφορετικά. Αφενός μεν η πολυπλοκότητα είναι της τάξης $O(|\mathcal{A}|^3)$ [44], αφετέρου υπάρχουν εμπειρικές ενδείξεις που δείχνουν το αντίθετο [8].

2.1.5 Άλλοι Αλγόριθμοι

Έχουν γίνει αρκετές μελέτες για την επιλογή του εύρους $u_{a,t}$, το οποίο συνήθως συναντάται στη βιβλιογραφία με το όνομα «συνάρτηση διόγκωσης» (padding function), κάθε μία από τις οποίες οδηγούν και σε έναν διαφορετικό αλγόριθμο. Χαρακτηριστικό παράδειγμα είναι ο αλγόριθμος MOSS (Minimax Optimal Stochastic Strategy) [2], ο οποίος χρησιμοποιεί πρότερη γνώση για το μέγεθος του ορίζοντα T και αντικαθιστά τον όρο $\log t$ της $u_{a,t}$ του UCB1 με $T/(|\mathcal{A}|N_{a,t})$ επιτυγχάνοντας λίγο καλύτερο άνω φράγμα μεταμέλειας.

Ο αλγόριθμος POKER (Price Of Knowledge And Estimated Reward) [63] χρησιμοποιεί τη μετρική της *αξίας της πληροφορίας* (value of information) που προκύπτει από κάθε δράση,

$$u_{a,t} = \mathbb{P}[\mu_a \geq \hat{\mu}_{\max} + \delta_\mu]\delta_\mu(T-t)$$



Σχήμα 2.3: Παράδειγμα Bayesian τύπου 2-armed bandit με Gaussian κατανομές επιβράβευσης. Αριστερά: Κατανομές των δύο δράσεων τη χρονική στιγμή της επιλογής. Η δράση a_2 έχει μεγαλύτερο άνω όριο αβεβαιότητας (κόκκινη κουκκίδα) και επιλέγεται παρόλο που έχει χαμηλότερη εκτιμώμενη επιβράβευση $\mu_2 < \mu_1$. Η επιβράβευση r_t που επιστρέφεται έχει χαμηλή τιμή (σημάδι «x»). Δεξιά: Η ύστερη συνάρτηση πυκνότητας πιθανότητας ανανεώνεται και η δράση που θα επιλεχθεί την επόμενη στιγμή είναι η δράση a_1.

όπου $\hat{\mu}_{\max} = \max_a\{\hat{\mu}_{a,t}\}$ ενώ ο όρος δ_μ είναι μία εκτίμηση για τη διαφορά μεταξύ της πραγματικής μέγιστης αξίας δράσης, και του μέγιστου δειγματικού μέσου ως $\mathbb{E}[\mu_{\max} - \hat{\mu}_{\max}]$. Για τον υπολογισμό του δ_μ ο αλγόριθμος δημιουργεί μία ταξινόμηση των εκτιμώμενων μέσων αξιών των δράσεων θεωρώντας μία συνάρτηση δεικτοδότησης $I : \{1, 2, \ldots, K\} \mapsto \mathcal{A}$ έτσι ώστε $\hat{\mu}_{I(1)} \geq \hat{\mu}_{I(2)} \geq \cdots \geq \hat{\mu}_{I(K)}$. Τότε

$$\delta_\mu = \frac{\hat{\mu}_{I(1)} - \hat{\mu}_{I(\sqrt{K})}}{\sqrt{K}}$$

ενώ χρησιμοποιώντας την εμπειρική τυπική απόκλιση $\hat{\sigma}_{a,t}$ των επιβραβεύσεων κάθε δράσης γίνεται η εκτίμηση

$$\mathbb{P}[\mu_a \geq \hat{\mu}_{\max} + \delta_\mu] \cong \int_{\hat{\mu}_{\max}+\delta_\mu}^{\infty} \mathcal{N}\left(r; \hat{\mu}_{a,t}, \frac{\hat{\sigma}_{a,t}}{\sqrt{N_{a,t}}}\right) dr$$

Παρόλο που το μέγεθος του ορίζοντα T είναι απαραίτητο για τους υπολογισμούς, ο POKER είναι αλγόριθμος *μηδενικής μεταμέλειας* (zero regret) αφού εγγυάται ότι η στρατηγική απόφασης θα συγκλίνει στη βέλτιστη καθώς το μέγεθος του ορίζοντα τείνει στο άπειρο.

Μία άλλη ενδιαφέρουσα τροποποίηση είναι ο αλγόριθμος KL-UCB [40] στον οποίο η συνάρτηση διόγκωσης προκύπτει με χρήση της απόκλισης Kullback-Leibler, ενώ μία διαφορετική κατηγορία προβλημάτων είναι τα Bayesian bandits, τα οποία εκμεταλλεύονται πρότερη γνώση

για της κατανομές πιθανότητας επιβράβευσης (μέχρι στιγμής δεν υπήρχαν υποθέσεις ή παραδοχές για τις κατανομές αυτές, εκτός του ότι είναι άνω φραγμένες). Εάν $p[\mathcal{R}]$ περιγράφει την πιθανότητα της κατανομής των επιβραβεύσεων και H_t είναι η ιστορία των δράσεων και των εκβάσεων, τότε υπολογίζεται η ύστερη πιθανότητα $p[\mathcal{R}|H_t]$. Η λογική αυτή ακολουθείται και από τους αλγορίθμους αισιοδοξίας στο μέτωπο της αβεβαιότητας (με τον αλγόριθμο Bayesian UCB) αλλά και από αλγορίθμους *πιθανοτικού ταιριάσματος* (με τον αλγόριθμο Thompson sampling).

Ώς παράδειγμα, υποθέτοντας Gaussian τύπου κατανομές πιθανότητας επιβραβεύσεων με $\mathcal{R}^a(r) = \mathcal{N}(r; \mu_a, \sigma_a^2)$, τότε μετά από ένα σύνολο δράσεων A_t και επιβραβεύσεων r_t η ύστερη πιθανότητα των παραμέτρων θα ακολουθεί τη σχέση

$$p[\mu_a, \sigma_a|H_t] \propto p[\mu_a, \sigma_a] \prod_{A_t=a} \mathcal{N}(r_t; \mu_a, \sigma_a^2)$$

και πλέον η τυπική απόκλιση μπορεί να χρησιοποιηθεί ως μέτρο για τη συνάρτηση διόγκωσης ώστε η απόφαση να λαμβάνεται με βάση την παρακάτω σχέση,

$$A_t = \underset{a\in\mathcal{A}}{\operatorname{argmax}} \left\{ \mu_a + c\frac{\sigma_a}{\sqrt{N_{a,t}}} \right\}$$

όπου c μία θετική σταθερά επιλογής. Ένα χαρακτηριστικό παράδειγμα για Bayesian bandit 2-επιλογών φαίνεται στο σχήμα 2.3.

Στις μεθόδους *ταιριάσματος πιθανότητας* η κάθε δράση επιλέγεται με βάση την πιθανότητα να είναι η βέλτιστη, μία λογική που ακολουθείται και στη soft-max Boltzmann μέθοδο που παρουσιάστηκε νωρίτερα. Αν η πολιτική $\pi(a)$ καθορίζει την κατανομή μάζας πιθανότητας στον χώρο των δράσεων, τότε $\pi(a) = \mathbb{P}[\mu_a > \mu_{a'}, \forall a' \neq a]$. Η ύστερη πιθανότητα $\pi(a|H_t)$ είναι δύσκολο να υπολογιστεί, ωστόσο ο αλγόριθμος Thompson sampling εφαρμόζει την λογική ότι,

$$\pi(a|H_t) = \mathbb{P}[\mu_a > \mu_{a'}, \forall a' \neq a|H_t] = \mathbb{E}_{\mathcal{R}|H_t}\left[\mathbb{I}\{a = \underset{a'\in\mathcal{A}}{\operatorname{argmax}}\, \mu_{a'}\}\right]$$

ενώ στη συνέχεια κάνει χρήση του κανόνα Bayes για να ανανεώσει τις ύστερες κατανομές $p[\mathcal{R}|H_t]$, δειγματοληπτεί τις παραμέτρους για κάθε κατανομή από αυτές, υπολογίζει την εκτιμώμενη αξία κάθε δράσης ως $\hat{\mu}_a = \mathbb{E}[\mathcal{R}^a]$ και εκτελεί τη δράση με τη μεγαλύτερη εκτιμώμενη τιμή. Ο αλγόριθμος Thompson sampling επιτυγχάνει το κάτω φράγμα των Lai και Robbins.

Κεφάλαιο 3

Μάθηση σε Δυναμικά Περιβάλλοντα μίας Κατάστασης

Συνεχίζοντας στα βήματα του προηγούμενου κεφαλαίου, είναι σαφές ότι η αναζήτηση της βέλτιστης δράσης μέσα από ένα σύνολο επιλογών έχει απασχολήσει στο παρελθόν αρκετούς διαφορετικούς τομείς, από τη θεωρία παιγνίων, τη μηχανική μάθηση και την επιχειρησιακή έρευνα. Προσφάτως το ενδιαφέρον αυτό έχει ανακάμψει καθώς αποτελεί τον θεμέλιο λίθο της πολλά υποσχόμενης έρευνας στην περιοχή της ενισχυτικής μηχανικής μάθησης. Έχουν αναπτυχθεί αρκετές τεχνικές, βασισμένες στη λογική δοκιμής και λάθους (trial and error), σε γενετικούς αλγορίθμους [35] έως και ελαχιστοποίησης εντροπίας [53], έχοντας οδηγήσει στη δημιουργία ορόσημων για την εκτίμηση της επίδοσης του κάθε αλγορίθμου. Οι στοχαστικές μηχανές επιβράβευσης πολλαπλών επιλογών (multi-armed bandits) που παρουσιάστηκαν στο προηγούμενο κεφάλαιο αποτελούν ένα τέτοιο ορόσημο κατά το οποίο μία μηχανή μάθησης εκτελεί μία δράση από ένα σύνολο διακριτών επιλογών σε κάθε βήμα, δέχεται κάποια επιβράβευση η οποία εμπεριέχει στοχαστικότητα ή θόρυβο κατά την παραγωγή της από το περιβάλλον και χρησιμοποιεί την πληροφορία αυτή ως ανάδραση στη βάση γνώσης της ώστε να επανακαθορίσει την πολιτική με την οποία λαμβάνει τις αποφάσεις, αντιμετωπίζοντας κάθε φορά το δίλημμα εξερεύνησης-αξιοποίησης [3].

Ένα καθημερινό παράδειγμα μπορεί να περιγράφει έναν εργαζόμενο ο οποίος φτάνει στο γραφείο του το πρωί και καλείται να επιλέξει μεταξύ δύο μηχανών παραγωγής καφέ. Οι μηχανές μπορούν να παράγουν το ίδιο είδος καφέ, αλλά η κάθε μία έχει διαφορετική αξιοπιστία ως προς τη γεύση, τον χρόνο παραγωγής και την πιθανότητα εμφάνισης ελαττώματος. Η γεύση και η ποσότητα καφέ που παράγει η κάθε μηχανή μπορεί να διαφέρει από μέρα σε μέρα, αλλά ας υποθέσουμε ότι η μία εκ των δύο είναι κατά μέσο όρο καλύτερη της άλλης. Ο άνθρωπος δεν αξιολογεί τις μηχανές από μία και μοναδική δοκιμή αλλά πραγματοποιεί πολλές δοκιμές μέχρι τελικά να καταλήξει στη μηχανή που προτιμά περισσότερο. Ας υποθέσουμε τώρα ότι η πρώτη από τις μηχανές καφέ επισκευάστηκε ή αναβαθμίστηκε χωρίς να το γνωρίζουν οι εργαζόμενοι, οι οποίοι ίσως ανακαλύψουν από μόνοι τους τη διαφορά. Εάν ο εργαζόμενος του παραδείγματος είχε ήδη συγκλίνει στο να επιλέγει την δεύτερη μηχανή και πλέον δεν εκτελεί εξερευνητικές δράσεις κατά την επιλογή του, τότε δεν θα ανακαλύψει ποτέ αυτή τη διαφορά.

Όπως μπορεί να γίνει αντιληπτό, η στοχαστική φύση των προβλημάτων που παρουσιάζονται στην καθημερινότητα δεν είναι σχεδόν ποτέ γνωστή (τουλάχιστον όχι πλήρως) και η στατική προσέγγιση περιορίζεται σε ένα μικρό σύνολο εφαρμογών. Όπως αναφέρουν οι Sutton και Barto [56],

> "Οι *περισσότερες μέθοδοι κάνουν σημαντικές υποθέσεις για τη στατική φύση του περιβάλλοντος, καθώς επίσης χρησιμοποιούν πρότερη γνώση για την οποία όμως δεν υπάρχει πρόσβαση σε ένα πραγματικό πρόβλημα...*"

Είναι λοιπόν σαφής η σημαντικότητα της ανάπτυξης προσαρμοστικών αλγορίθμων λήψης αποφάσεων. Υπάρχει μία πληθώρα αλγορίθμων που έχουν ήδη αναπτυχθεί και επιτυγχάνουν καλή συμπεριφορά όσον αφορά τη συνολική μεταμέλεια των δράσεων. Σε αυτή την εργασία θα αναφέρουμε ένα σημαντικό υποσύνολο των σημαντικότερων απο αυτούς, βάσει της αναλυτικής και εμπειρικής τους επίδοσης. Στο [19] αποδεικνύεται ότι οι αλγόριθμοι *Discounted-UCB (D-UCB)* (ο οποίος πρωτοπαρουσιάστηκε στο [34]) και *Sliding Window-UCB (SW-UCB)* επιτυγχάνουν ικανοποιητική ασυμπτωτική συμπεριφορά της μεταμέλειας. Για να εγγυώνται ωστόσο αυτή τη συμπεριφορά, χρειάζεται η κατάλληλη επιλογή των παραμέτρων τους η οποία προϋποθέτει κάποια πρότερη γνώση για τη δυναμική του περιβάλλοντος - το ρυθμό με τον οποίο αλλάζει η βέλτιστη επιλογή καθώς και το μέγεθος του ορίζοντα. Ο αλγόριθμος *Kalman Filter-Multi Armed Bandit (KF-MANB)* όπως αναφέρεται στο [21], είναι μία Bayesian προσέγγιση και επιτυγχάνει αρκετά καλά αποτελέσματα με έναν κομψό και διαισθητικά προσιτό μαθηματικό φορμαλισμό. Τέλος, ο αλγόριθμος *Adapt-EvE* [22], ο οποίος επέδειξε πολύ καλά εμπειρικά αποτελέσματα στο σύνολο προβλημάτων PASCAL-EvE 2006, χρησιμοποιεί έναν ανιχνευτή αλλαγής κατάστασης (change-point detector) βασισμένο σε στατιστική Page-Hinkley σε συνδυασμό με τον αλγόριθμο UCB-Tuned για τη λήψη αποφάσεων.

3.1 Κατηγορίες Μη-Στάσιμων Περιβάλλοντων

Υποθέτουμε ότι ένα σύστημα αλληλεπιδράσεων εισόδου-εξόδου, μπορεί να περιγραφεί σε κάθε στιγμιότυπό του με ένα πλήθος καθορισμένων στατιστικών νόμων, κάθε ένας από τους οποίους μπορεί επίσης να μοντελοποιηθεί με μία κατανομή πιθανότητας. Υποθέτουμε ότι η κατανομή πιθανότητας είναι παραμετρική, ή μπορεί να περιγραφεί ικανοποιητικά από ένα σύνολο παραμέτρων θ_t κάθε χρονική στιγμή t. Οι παράμετροι θ_t μπορούν να αλλάζουν από τη μία χρονική στιγμή στην άλλη, έτσι ώστε $\theta_{t+1} \neq \theta_t$. Περιβάλλοντα στα οποία οι αλλαγές μεταξύ διαδοχικών στιγμών είναι μικρές, δηλαδή $\|\theta_{t+1} - \theta_t\| \leq \epsilon$, όπου ϵ είναι μια μικρή σταθερά λέγονται *ομαλά* (drifting), ενώ περιβάλλοντα για τα οποία, $\exists t : \|\theta_{t+1} - \theta_t\| \geq \delta$, όπου δ είναι μια ικανοποιητικά μεγάλη τιμή, λέγονται *διακοπτόμενα* (switching/abrupt), ενώ τα ϵ, δ μπορούν να χρησιμοποιηθούν και ως αναφορές για τον χαρακτηρισμό του περιβάλλοντος.

Στα ομαλά δυναμικά περιβάλλοντα μπορεί κανείς να χρησιμοποιήσει την ιστορία των τιμών των παραμέτρων ώστε να προβεί σε μελλοντική πρόβλεψη. Ακόμα και αν δεν υπάρχει κάποια δυναμική, όπως για παράδειγμα στην περίπτωση που το θ αντιστοιχεί σε ένα σημείο

που εκτελεί Brownian κίνηση εντός του παραμετρικού χώρου, μπορεί οι βραχέως χρόνου εκτιμήσεις να μην είναι σωστές αλλά το σφάλμα πρόβλεψης θα είναι μικρό. Στα διακοπτόμενα περιβάλλοντα υποθέτουμε ότι οι αλλαγές μπορεί να είναι μερικώς προβλέψιμες α) ως προς το πότε θα συμβούν και β) ως προς το ποιες θα είναι οι νέες τιμές θ_{t+1}, μόνο σε ειδικές κατηγορίες (όπως σε περιοδικές εναλλαγές βέλτιστων δράσεων). Παρ΄ όλα αυτά υπάρχει ένα μεγάλο σύνολο περιπτώσεων (όπως για παράδειγμα όταν οι αλλαγές καθορίζονται από μια γεννήτρια τυχαίων αριθμών) όπου κάτι τέτοιο δεν είναι εφικτό. Τα προβλήματα που μας ενδιαφέρουν στη συγκεκριμένη εργασία είναι του δεύτερου τύπου, προβλήματα δηλαδή στα οποία οι αλλαγές δεν είναι ανιχνεύσιμες χρονικά και χωρικά. Διατυπώνοντας σε μαθηματική μορφή το συγκεκριμένο σκεπτικό, εάν $c_t \in \{0, 1\}$ είναι μία δυαδική μεταβλητή που καθορίζει την ύπαρξη ή όχι αλλαγής των παραμέτρων τη χρονική στιγμή t, τότε για τα προβλήματα αυτά θα ισχύει ότι $p(\theta_{t+1}|\theta_t, c_t) = p(\theta_{t+1}|c_t)$.

3.2 Αλγόριθμοι για Μη-Στάσιμα MABs

Αν και υπάρχουν αρκετοί αλγόριθμοι για την προσέγγιση του προβλήματος MAB σε μη στατικά περιβάλλοντα, στα επόμενα τμήματα ακολουθεί μία επιλογή από τους βασικότερους. Σε κάθε περίπτωση, η βέλτιστη πολιτική ενός αλγορίθμου που επιτυγχάνει λογαριθμική μεταμέλεια σε ένα στάσιμο περιβάλλον, δεν μπορεί να έχει καλύτερη ασυμπτωτική συμπεριφορά από $O(T/\log T)$. Ακολουθώντας το σκεπτικό των Garivier και Moulines από το [19], έστω $\mathcal{Q}$ μία συνάρτηση πυκνότητας πιθανότητας επιβραβεύσεων τέτοια ώστε η μέση εκτιμώμενη επιβράβευση $\nu = \mathbb{E}_{\mathcal{Q}}[R]$ να είναι μεγαλύτερη από τη μέγιστη εκτιμώμενη επιβράβευση των κατανομών $\mathcal{R}^a$, ή απλούστερα $\nu > \mu_{I(1)}$, όπου $I(\cdot)$ η συνάρτηση δεικτοδότησης όπως αναφέρθηκε στο προηγούμενο κεφάλαιο. Έστω $\delta = \nu - \mu_{I(1)}$ και Δ_G το μέγιστο χάσμα δράσεων. Έστω επίσης ότι $b = KL(\mathcal{R}^{I(K)}||\mathcal{Q})$ η Kullback-Leibler απόκλιση μεταξύ της κατανομής της χειρότερης δράσης (με βάση τη μέση επιβράβευση) και της κατανομής $\mathcal{Q}$, και ότι $\mathbb{E}_\pi[\cdot]$ είναι η εκτίμηση υπο την πολιτική π. Για τη μοντελοποίηση της μη στατικότητας, οι χρονικές στιγμές $\{1, 2, \ldots, T\}$ χωρίζονται σε τμήματα μεγέθους τ κάθε ένα από τα οποία δεικτοδοτείται με έναν αύξοντα δείκτη j, ώστε $1 \leq j \leq M = \lfloor \frac{T}{\tau} \rfloor$. Με $\mathbb{P}_\pi$ συμβολίζεται η συνολική κατανομή των κερδών από την αρχή του χρόνου έως το τέλος του ορίζοντα υπό την πολιτική π. Εάν κατά την χρονική περίοδο j η κατανομή επιβραβεύσεων της χειρότερης δράσης $\mathcal{R}^{I(K)}$ αλλάξει σε ν, έτσι ώστε η χειρότερη δράση να γίνει η βέλτιστη για το j-οστό διάστημα, η κατανομή κερδών του περιβάλλοντος αυτού συμβολίζεται με $\mathbb{P}^j_\pi$ ενώ $\mathbb{E}^j_\pi[\cdot]$ θα είναι η εκτίμηση υπο την πολιτική π όταν το περιβάλλον υπόκειται σε αυτή την απότομη αλλαγή. Εάν η επιλογή του δείκτη j είναι μεν στοχαστική αλλά γίνεται με ισοπίθανο τρόπο για όλα τα διαστήματα, τότε $\mathbb{P}^*_\pi$ θα είναι η αντίστοιχη κατανομή των επιβραβεύσεων και $\mathbb{E}^*_\pi[\cdot]$ η εκτίμηση υπό την πολιτική π σε αυτά τα διακοπτόμενου τύπου μη στάσιμα περιβάλλοντα, όπου

$$\mathbb{E}^*_\pi[\cdot] = \frac{1}{M} \sum_{j=1}^{M} \mathbb{E}^j_\pi[\cdot]$$

Το κάτω φράγμα της συνολικής εκτιμώμενης μεταμέλειας $\mathbb{E}^*_\pi[L_T]$ που μπορεί να επιτύχει μία πολιτική π στα διακοπτόμενα μη-στατικά περιβάλλοντα $\mathbb{P}^*_\pi$ θα εξαρτάται από την συνολική εκτιμώμενη μεταμέλεια $\mathbb{E}_\pi[L_T]$ της πολιτικής αυτής στα στάσιμα περιβάλλοντα $\mathbb{P}_\pi$, με τρόπο που ορίζεται από το παρακάτω θεώρημα.

Θεώρημα 3.4. *(Θεώρημα 13 από [19]) Για κάθε πολιτική π και ορίζοντα T τέτοια ώστε $64/(9a) \leq \mathbb{E}_\pi[N_{I(K),T}] \leq T/(4a)$, ισχύει ότι*

$$\mathbb{E}^*_\pi[L_T] \geq C(\mu)\frac{T}{\mathbb{E}_\pi[L_T]},\ \text{όπου } C(\mu) = \Delta_G\frac{32\delta}{27b}$$

Όμως $\max\{a, n/a\} \geq \sqrt{n}$, για κάθε $a, n > 0$ (κάτι που αποδεικνύεται και εύκολα με δύο υποθέσεις), οπότε

$$\max\{\mathbb{E}_\pi[L_T], \mathbb{E}_{\pi^\star}[L_T]\} \geq \max\left\{\mathbb{E}_\pi[L_T], \frac{C(\mu)T}{\mathbb{E}_\pi[L_T]}\right\} \geq \sqrt{C(\mu)T} \tag{3.1}$$

δηλαδή καμία πολιτική σε μη-στατικά περιβάλλοντα δεν μπορεί να επιτύχει μεταμέλεια καλύτερη από $L_T = O(\sqrt{T})$. Επίσης προκύπτει ότι για κάθε πολιτική π που επιτυγχάνει τη βέλτιστη ασυμπτωτική λογαριθμική συμπεριφορά σε ένα στατικό περιβάλλον, δηλαδή $\mathbb{E}_\pi[L_T] = \Theta(\log T)$, όπως για τους UCB αλγορίθμους, θα ισχύει $\mathbb{E}^*_\pi[L_T] = O(T/\log T)$

Για την παρουσίαση των αλγορίθμων και των βασικότερων χαρακτηριστικών, υποθέτουμε ότι οι επιβραβεύσεις που επιστρέφει το περιβάλλον είναι φραγμένες στο $[0, B]$ ενώ με Y_T συμβολίζεται ο αριθμός των διακοπτόμενων αλλαγών που συμβαίνουν στο περιβάλλον για ορίζοντα T.

3.2.1 Αλγόριθμος Discounted UCB

Όπως όλες οι περισσότερες πολιτικές της οικογένειας UCB, ο D-UCB αλγόριθμος που παρουσιάστηκε στο [34] και αναλύεται στο [19], χρησιμοποιεί τον αριθμό $N_{a,t}$ των φορών που έχει επιλεχθεί κάθε δράση καθώς και τους εμπειρικούς μέσους των επιβραβεύσεων $\hat{\mu}_{a,t}$. Όσο περισσότερο έχει επιλεχθεί μία δράση, τόσο μικρότερη θα είναι και η τιμή της αντίστοιχης συνάρτησης διόγκωσης $u_{a,t}$ και τόσο μικρότερη αλλαγή στην εκτιμώμενη μέση επιβράβευση $\mu_{a,t}$ θα επιφέρει ένα νέο δείγμα, ακόμα και αν αυτό απέχει αρκετά από την τρέχουσα εκτιμώμενη τιμή. Έτσι, δημιουργείται το πρόβλημα της *αδράνειας μάθησης* και της χαμηλής προσαρμοστικότητας. Η ιδέα που εφαρμόζει ο D-UCB για τη μείωση της αδράνειας μάθησης είναι να μειώνει εκθετικά τη σημαντικότητα των παρελθοντικών εμπειριών, με εκθετική μείωση των παρελθοντικών επιβραβεύσεων αλλά και των φορών που έχει επιλεγεί η κάθε δράση.

Χρησιμοποιώντας έναν *παράγοντα υποβάθμισης* γ, με $\gamma \in (0,1)$, ο τρέχων σταθμισμένος μέσος $\tilde{\mu}_{a,t}$ θα είναι

$$\tilde{\mu}_{a,t} = \frac{1}{\tilde{N}_{a,t}} \sum_{s=1}^{t} \gamma^{t-s} r_s \mathbb{I}\{A_s = a\}$$

όπου $\tilde{N}_{a,t} = \sum_{s=1}^{t} \gamma^{t-s} \mathbb{I}\{A_s = a\}$, ενω η νέα τροποποιημένη συνάρτηση διόγκωσης $u_{a,t}$ για κάθε δράση a θα είναι

$$u_{a,t} = 2B \sqrt{\frac{\xi \log \sum_{a' \in \mathcal{A}} \tilde{N}_{a',t}}{\tilde{N}_{a,t}}}$$

όπου ξ μία παράμετρος επιλογής, ενώ μπορεί κανείς να χρησιμοποιήσει την ίδια λογική του UCB-Tuned που αναφέραμε στην στατική περίπτωση του προηγούμενου κεφαλαίου ώστε να καταλήξει στον αλγόριθμο D-UCB-Tuned.

Αποδεικνύεται ότι εάν $\check{N}_{a,T}$ συμβολίζει το πόσες φορές που επιλέχθηκε η δράση a χωρίς να είναι η βέλτιστη, δηλαδή $\check{N}_{a,t} = \sum_{t=1}^{T} \mathbb{I}\{A_t = a \neq a_t^\star\}$, τότε για παράμετρο $\xi > 1/2$ και $\gamma \in (0,1)$, ο εκτιμώμενος αριθμός των εσφαλμένων δράσεων $\mathbb{E}_\gamma[\check{N}_{a,T}]$ είναι άνω φραγμένος, κάτι που στη συνέχεια οδηγεί στην εύρεση της ασυμπτωτικής συμπεριφοράς της μεταμέλειας του D-UCB. Συγκεκριμένα, υποθέτοντας γνώση του συνολικού αριθμού των διακοπτόμενων αλλαγών του περιβάλλοντος Y_T για ορίζοντα T, τότε εάν ο παράγοντας υποβάθμισης γ επιλεχθεί ώστε $\gamma = 1 - (4B)^{-1}\sqrt{Y_T/T}$ προκύπτει ότι

$$\mathbb{E}_\gamma \left[\check{N}_{a,T}\right] = O\left(\sqrt{TY_T} \log T\right)$$

και όπως προκύπτει από την εξίσωση (3.1) του θεωρήματος 3.4, ο αλγόριθμος D-UCB επιτυγχάνει υποβέλτιστη συμπεριφορά κατά έναν παράγοντα $\log T$.

3.2.2 Αλγόριθμος Sliding Window UCB

Μία εναλλακτική προσέγγιση που προτείνεται και αναλύεται επίσης στο [19], είναι ο αλγόριθμος Sliding Window UCB (SW-UCB) όπου αντί της χρήσης ενός παράγοντα υποβάθμισης για τον υπολογισμό του τρέχοντα δειγματικού μέσου $\tilde{\mu}_{a,t}$ κάθε δράσης, η ιδέα είναι να χρησιμοποιηθεί ένα χρονικό παράθυρο μεγέθους τ. Έτσι,

$$\tilde{\mu}_{a,t} = \frac{1}{\tilde{N}_{a,t}} \sum_{s=t-\tau+1}^{t} r_s \mathbb{I}\{A_s = a\}, \text{ όπου } \tilde{N}_{a,t} = \sum_{s=t-\tau+1}^{t} \mathbb{I}\{A_s = a\}$$

θεωρώντας[2] μηδενικές επιβραβεύσεις και μηδενικούς δείκτες για $t < \tau$. Η συνάρτηση διόγκωσης θα είναι

$$u_{a,t} = B\sqrt{\frac{\xi \log(t \wedge \tau)}{\tilde{N}_{a,t}}}$$

όπου $t \wedge \tau \equiv \min\{t, \tau\}$ και ξ επίσης μία παράμετρος επιλογής. Αν $\xi > 1/2$ και τα Y_T, T είναι γνωστά, τότε επιλέγοντας χρονικό παράθυρο $\tau = 2B\sqrt{T \log T / Y_T}$

$$\mathbb{E}_\tau \left[\check{N}_{a,t}\right] = O(\sqrt{T Y_T \log T})$$

δηλαδή είναι ελαφρώς καλύτερος από τον D-UCB καθώς δεν επιτυγχάνει τη βέλτιστη μεταμέλεια $O(\sqrt{T})$ κατά έναν παράγοντα $\sqrt{\log T}$.

3.2.3 Αλγόριθμος Adapt-EvE

Μία άλλη προσέγγιση είναι ο Αλγόριθμος Adapt-EvE, ο οποίος έχει καλά εμπειρικά αποτελέσματα σε ένα σύνολο επιλεγμένων προβλημάτων [22]. Στην ουσία το κίνητρο για την ανάπτυξη του Adapt-EvE ήταν η βελτιστοποίηση λήψης αποφάσεων σε πραγματικά προβλήματα που παρουσιάζονται σε ιστοσελίδες, με σκοπό να παρουσιάζονται στον χρήστη οι κατάλληλες ειδήσεις για τις οποίες θα υπήρχε ενδιαφέρον για περαιτέρω ανάγνωση. Το πρόβλημα μοντελοποιήθηκε ως ένα multi-armed bandit, όπου η κάθε είδηση αντιστοιχούσε σε μία δράση και οι επιβραβεύσεις ήταν τύπου Bernoulli, ενώ $r_t = 1$ σήμαινε πως ο χρήστης επέλεξε προς ανάγνωση την είδηση και $r_t = 0$ σήμαινε πως δεν έδειξε ενδιαφέρον. Ο Adapt-EvE έπρεπε να είναι προσαρμοστικός ώστε να ανιχνεύει πιθανές αλλαγές στις προτιμήσεις του χρήστη και να τροποποιεί κατάλληλα τη στρατηγική εμφάνισης ειδήσεων.

Για τον σκοπό αυτό χρησιμοποιεί έναν ανιχνευτή διακοπτόμενων αλλαγών του περιβάλλοντος (change-point detector) βασισμένο σε στατιστικό έλεγχο τύπου Page-Hinkley. Η

[2] Σε μία από τις εκδόσεις των σχετικών δημοσιεύσεων των Gariver, Moulines υπάρχει ένα τυπογραφικό λάθος καθώς χρησιμοποιούν την ίδια έκδοση για το $\tilde{N}_{a,t}$ όπως στον D-UCB το οποίο όμως δεν μπορεί να είναι σωστό για τους λόγους που αναφέρεται στο [9]. Παρ΄ όλα αυτά η προσέγγιση στο [9] δεν φαίνεται σωστή. Η έκδοση που παρουσιάζεται εδώ είναι αυτή που αναγράφεται στη τελευταία διορθωμένη δημοσίευση, ή αντίστοιχα η ίδια που αναφέρεται και στη διδακτορική διατριβή του Mellor, University of Manchester [42]

υπόθεση H_0 είναι ότι το περιβάλλον είναι στατικό και μπορεί να περιγραφεί από ένα σύνολο στατιστικών παραμέτρων θ. Παράλληλα με την αλληλεπίδραση της μηχανής μάθησης με το περιβάλλον και την παρατήρηση των εκβάσεων μετά από κάθε δράση, η οποία λαμβάνεται με έναν αλγόριθμο UCB-Tuned (ή D-UCB-Tuned), γίνεται και ο έλεγχος για την απόρριψη ή όχι της H_0. Γι΄ αυτόν το σκοπό, εάν $\hat{r}_t$ συμβολίζει τον απλό εμπειρικό μέσο των συνολικών επιβραβεύσεων, υπολογίζει τους όρους

$$m_T = \sum_{t=1}^{T}(r_t - \hat{r}_t + \delta) \ , \ M_T = \max_t\{m_t\} \ , \ H_T = M_T - m_T - \lambda$$

όπου δ, λ παράμετροι επιλογής που ελέγχουν την ευαισθησία του τεστ. Εάν $H_T > 0$ τότε θεωρεί ότι ανιχνεύθηκε αλλαγή των στατιστικών του περιβάλλοντος. Σε αυτή την περίπτωση, μία πρώτη προσέγγιση θα ήταν η επανεκκίνηση του UCB-Tuned, διαγράφοντας την ιστορία και αρχικοποιώντας όλες τις τιμές $\hat{\mu}_{a,t}$. Στις περιπτώσεις όμως εσφαλμένης αναγνώρισης της διακοπής (σφάλματα τύπου Ι) η απώλεια της πληροφορίας μπορεί να είναι μεγάλη και να επιφέρει σημαντική μείωση στην επίδοση. Γι΄ αυτό το λόγο ο Adapt-EvE έχει μία πιο ενδιαφέρουσα προσέγγιση. Κατά την απόρριψη της H_0 αρχικοποιεί έναν δεύτερο UCB-Tuned αλγόριθμο στο ίδιο επίπεδο (επίπεδο 1) με τον αρχικό, και έναν τρίτο μετα-αλγόριθμο UCB (είτε κάποιον άλλο από την οικογένεια UCB) σε ένα επίπεδο πάνω από τους υπόλοιπους (επίπεδο 2). Ο μετα-αλγόριθμος του επιπέδου 2 επιλέγει ποιος από τους δύο αλγορίθμους του επιπέδου 1 (ο παλιός ή ο νέος) θα λάβει την τελική απόφαση, ανανεώνοντας τη γνώση του μετά από κάθε έκβαση. Μετά την πάροδο κάποιας προκαθορισμένης χρονικής περιόδου, διατηρείται μόνο ο αλγόριθμος επιπέδου 1 για τον οποίο αντιστοιχεί η μεγαλύτερη μέση εκτιμώμενη επιβράβευση βάσει των παρατηρήσεων του μετα-αλγορίθμου στο επίπεδο 2.

Εάν και η προσέγγιση αυτή είναι αρκετά ενδιαφέρουσα, ο Page-Hinkley έλεγχος κατά τη χρονική διάρκεια που ο μετα-αλγόριθμος είναι ενεργός είναι απενεργοποιημένος και κάθε νέα αλλαγή που μπορεί να συμβεί σε αυτό το διάστημα δεν ανιχνεύεται. Παρόλο που ο Adapt-EvE έχει πολύ καλά εμπειρικά αποτελέσματα σε πολλές περιπτώσεις, υπάρχει ανάγκη για πολύ προσεκτική ρύθμιση των παραμέτρων οι οποίες μπορεί να διαφέρουν αρκετά από περιβάλλον σε περιβάλλον.

3.2.4 Αλγόριθμος KF-MANB

Ο αλγόριθμος αυτός [21] προτείνει τη χρήση φίλτρων Kalman και κατά κύριο λόγο χρησιμοποιείται όταν οι επιβραβεύσεις ακολουθούν κανονική κατανομή. Παρ΄ όλα αυτά, λόγω του κεντρικού οριακού θεωρήματος το βραχυπρόθεσμο άθροισμα των επιβραβεύσεων θα ακολουθεί σχεδόν-κανονική κατανομή, συνεπώς δεν θεωρείται κακή επιλογή ακόμα και σε περιπτώσεις Bernoulli MABs (όπως θα δούμε στη συνέχεια επιτυγχάνει καλύτερα εμπειρικά αποτελέσματα από τους υπόλοιπους σε ένα σύνολο επιλεγμένων προβλημάτων). Για τη θεωρητική προσέγγιση θα χρησιμοποιήσουμε κάποια στοιχεία από το [16] σε Bayesian εκτίμηση παραμέτρων,

στην περίπτωση που έχουμε κανονικές κατανομές.

Υποθέτουμε ότι η μεταβλητή r ακολουθεί κανονική κατανομή $p(r|\mu) \sim \mathcal{N}(\mu, \sigma^2)$, όπου ο μέσος μ είναι άγνωστος. Ωστόσο μπορούμε να υποθέσουμε πως $p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Εάν $r_1, r_2, \ldots, r_n$ είναι ανεξάρτητα δείγματα από αυτή την κατανομή, με $\mathcal{D} = \{r_1, \ldots, r_n\}$ το σύνολο δειγμάτων, τότε από τον κανόνα του Bayes προκύπτει

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)d\mu} = \alpha \prod_{k=1}^{n} p(r_k|\mu)p(\mu)$$

όπου α είναι ένας παράγοντας κανονικοποίησης. Αντικαθιστώντας τις κατανομές για τα $p(r_k|\mu)$ και $p(\mu)$ στην παραπάνω σχέση, και ομαδοποιώντας κατάλληλα τους όρους εντός του εκθετικού, καταλήγουμε στο ότι

$$p(\mu|\mathcal{D}) = \alpha' \exp\left\{-\frac{1}{2}\left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} r_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right\}$$

όπου η σταθερά κανονικοποίησης α' έχει απορροφήσει και τους όρους που δεν εξαρτώνται από το μ. Υποθέτοντας ότι $p(\mu|\mathcal{D}) = \mathcal{N}(\mu_n, \sigma_n^2)$ και εξισώνοντας τους αντίστοιχους όρους, προκύπτει ότι

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{r}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \quad , \quad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} \tag{3.2}$$

δηλαδή ο νέος εκτιμώμενος μέσος για το μ μετά από n δείγματα θα είναι ένας κυρτός συνδυασμός του δειγματικού μέσου $\bar{r}_n$ και της προηγούμενης εκτίμησης μ_0, ενώ η αβεβαιότητα για τον μέσο αυτό θα μειώνεται διαρκώς.

Η παραπάνω Bayesian προσέγγιση μπορεί να ακολουθηθεί στην περίπτωση στατικών στοχαστικών MABs. Όμως στην περίπτωση ενός δυναμικού περιβάλλοντος κάτι τέτοιο δεν θα ήταν σωστό για δύο λόγους. Αφενός ως προς τον υπολογισμό της καλύτερης εκτίμησης για τον μέσο μ καθώς τα νέα δείγματα ίσως είναι πιο σημαντικά, και αφετέρου ως προς τον υπολογισμό της αβεβαιότητας η οποία θα πρέπει να αυξάνεται όταν τα δείγματα αυτά δεν ακολουθούν τις εκτιμώμενες στατιστικές προβλέψεις. Ο αλγόριθμος KF-MANB είναι μία απλή προσέγγιση χρήσης φίλτρου Kalman για τον παραπάνω σκοπό. Θεωρεί τον θόρυβο σ_{ob}^2 ο οποίος σχετίζεται με τη στοχαστικότητα των παρατηρήσεων (αντίστοιχος του σ^2 στην προηγούμενη προσέγγιση) και έναν επιπλέον *θόρυβο μετάβασης* σ_{tr}^2. Για τη λήψη απόφασης χρησιμοποιεί τις εκτιμήσεις $\hat{\mu}_{a,t}$ για τη μέση επιβράβευση κάθε δράσης, καθώς και τις εκτιμήσεις $\hat{\sigma}_{a,t}$ για τις

αντίστοιχες τυπικές αποκλίσεις. Με τις τρέχουσες εκτιμήσεις αυτές λαμβάνεται ένα τυχαίο δείγμα $x_{a,t}$ από κάθε κατανομή $\mathcal{N}(\hat{\mu}_{a,t}, \hat{\sigma}^2_{a,t})$ και η δράση επιλογής γίνεται ως

$$A_t = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, x_{a,t}$$

Θεωρούμε ότι η δράση A_t εκτελείται και παρατηρείται η επιβράβευση r_t. Οι εκτιμήσεις για όλες τις υπόλοιπες δράσεις $a \in \mathcal{A} \backslash A_t$ ανανεώνονται, με βάση τις παρακάτω σχέσεις

$$\hat{\mu}_{a,t+1} = \hat{\mu}_{a,t} \quad , \quad \hat{\sigma}^2_{a,t+1} = \hat{\sigma}^2_{a,t} + \sigma^2_{tr} \tag{3.3}$$

δηλαδή οι μέσοι παραμένουν αμετάβλητοι ενώ οι διασπορές αυξάνονται με την προσθήκη του θορύβου μετάδοσης. Για την επιλεγμένη δράση $a \equiv A_t$ οι νέες εκτιμήσεις θα είναι

$$\hat{\mu}_{a,t+1} = \frac{\hat{\sigma}^2_{a,t} + \sigma^2_{tr}}{\hat{\sigma}^2_{a,t} + \sigma^2_{tr} + \sigma^2_{ob}} r_t + \frac{\sigma^2_{ob}}{\hat{\sigma}^2_{a,t} + \sigma^2_{tr} + \sigma^2_{ob}} \hat{\mu}_{a,t} \quad , \quad \hat{\sigma}^2_{a,t+1} = \frac{(\hat{\sigma}^2_{a,t} + \sigma^2_{tr})\sigma^2_{ob}}{\hat{\sigma}^2_{a,t} + \sigma^2_{tr} + \sigma^2_{ob}} \tag{3.4}$$

και η διαδικασία συνεχίζεται κυκλικά. Οι αρχικές εκτιμήσεις $\hat{\mu}_{a,0}$ των δράσεων αρχικοποιούνται (π.χ με αισιόδοξη αρχικοποίηση) ενώ για τις αρχικές τιμές των τυπικών αποκλίσεων $\hat{\sigma}_{a,0}$ είναι εύλογο να επιλεχθεί μία τιμή κοντά στο εύρος B των επιβραβεύσεων.

Είναι εμφανές ότι ο παραπάνω φορμαλισμός προκύπτει από τη Bayesian προσέγγιση για την εκτίμηση παραμέτρων χρησιμοποιώντας $n = 1$ και αντικαθιστώντας τον θόρυβο σ^2_0 με $\hat{\sigma}^2_{a,t} + \sigma^2_{tr}$ για κάθε δράση.

Κεφάλαιο 4

Βιολογικά Εμπνευσμένη Μετα-Μάθηση

4.1 Μετα-Μάθηση και Νευροτροποποίηση

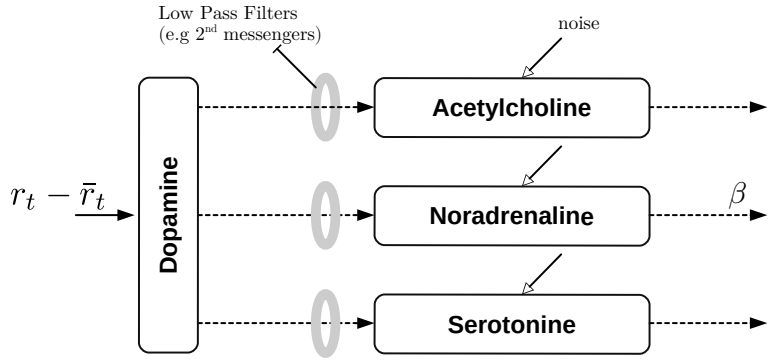
Στα πεδία της ρομποτικής μάθησης [31, 28, 6] και των υπολογιστικών νευροεπιστημών μάθησης και λήψης αποφάσεων [15, 12, 27, 58], αρκετές έρευνες αντιμετωπίζουν το δίλημμα εξερεύνησης-αξιοποίησης με τη χρήση της συνάρτησης soft-max Boltzmann, χρησιμοποιώντας πολλές φορές εξερευνητικές επιβραβεύσεις για κάποιες δράσεις (exploration bonuses) [13, 18]. Παρόλο που η προσέγγιση αυτή είναι πολλά υποσχόμενη λόγω των πολύ καλών εμπειρικών αποτελεσμάτων, υπάρχουν πολλοί περιορισμοί που μπορεί να σκεφτεί κανείς. Αρχικά, με τη χρήση μίας σταθερής αμετάβλητης παραμέτρου αντίστροφης θερμοκρασίας β (με την οποία ελέγχονται τα επίπεδα εξερευνητικών δράσεων μέσω της soft-max συνάρτησης) ή με τη συνεχή αύξηση αυτής της παραμέτρου, το χρονικό εύρος των εξερευνητικών δράσεων δεν είναι αρκετό για την ανίχνευση των αλλαγών της βέλτιστης δράσης στα δυναμικά περιβάλλοντα. Στο [50] προτάθηκε μία βιολογικής φύσης προσέγγιση για την προσαρμογή των μετα-παραμέτρων ενός αλγορίθμου ενισχυτικής μάθησης, χρησιμοποιώντας ως ανάδραση στοιχεία που σχετίζονται με τη φασική και τονική πυροδότηση των νευρώνων ντοπαμίνης. Συγκεκριμένα στο [15] αναφέρεται η παρακάτω σχέση κάποιων νευροδιαβιβαστών με τους μηχανισμούς μάθησης.

Ντοπαμίνη: Αναπαριστά το γενικό σήμα ελέγχου που καθορίζει την ενίσχυση ή όχι της σημαντικότητας των εμπειριών (άρα και της μάθησης).

Σεροτονίνη: Ελέχει τη σημαντικότητα μεταξύ των βραχέως χρόνου και μακρέως χρόνου προβλέψεων επιβράβευσης.

Νοραδρεναλίνη: Ελέγχει την αναλογία κατά την επιλογή μεταξύ εξερευνητικών δράσεων και αξιοποιητικών δράσεων.

Ακετυλοχολίνη: Ελέγχει την ισορροπία μεταξύ της αποθήκευσης των εμπειριών στη μνήμη, ή την ανανέωση της μνήμης.



Σχήμα 4.1: Πιθανή εξάρτηση νευροδιαβιβαστών από τα σήματα επιβράβευσης. Το σήμα που ρυθμίζει την πυροδότηση των νευρώνων ντοπαμίνης εκτιμάται ως η διαφορά μεταξύ της στιγμιαίας επιβράβευσης και της μέσης εκτιμώμενης επιβράβευσης. Το σήμα που καταφθάνει στους νευρώνες νοραδρεναλίνης και συσχετίζεται με τα επίπεδα εξερεύνησης, δέχεται προσεγγιστικά ένα βαθυπερατό φιλτράρισμα λόγω των βιολογικών διεργασιών που μεσολαβούν.

Ωστόσο, όπως αναφέρεται στο [15], το σήμα που ρυθμίζει την πυροδότηση των νευρώνων ντοπαμίνης μπορεί να σχετίζεται με τη διαφορά της στιγμιαίας επιβράβευσης r_t και της μέσης εκτιμώμενης επιβράβευσης $\bar{r}_t$. Το σήμα αυτό μεταφέρεται στους επόμενους νευρο-ρυθμιστικούς νευρώνες μέσω βιολογικών μηχανισμών που ενδεχομένως εμπεριέχουν κάποιου τύπου βαθυπερατό φιλτράρισμα (π.χ λόγω φαινομένων διάχυσης ή την ύπαρξη 2ης τάξης αγγελιοφόρων που μεταφέρουν την ντοπαμίνη). Έτσι, το σήμα που αναμένεται να ρυθμίζει την πυροδότηση των νευρώνων νοραδρελανίνης (ή νορεπινεφρίνης), είναι η διαφορά μεταξύ της μέσης επιβράβευσης βραχέως χρόνου $\bar{r}_t$ και της μέσης επιβράβευσης δεύτερης τάξης (ή μακρέως χρόνου) $\bar{\bar{r}}_t$. Κάτι τέτοιο συμβαδίζει μερικώς και από την παλαιότερη παρατήρηση των Solomon και Corbit [54] στον τομέα της ψυχολογίας για την επίδραση των μέσων θετικών επιβραβεύσεων και των μακρέως χρόνου αρνητικών εμπειριών στα κίνητρα και το συναίσθημα των ανθρώπων. Για τον λόγο αυτό στην ίδια έρευνα προτάθηκε η ρύθμιση της αναλογίας εξερευνητικών και αξιοποιητικών δράσεων από το σήμα $\bar{r}_t - \bar{\bar{r}}_t$, σε αλγορίθμους ενισχυτικής μάθησης (Σχήμα 4.1).

4.2 Απλός Αλγόριθμος Μετα-Μάθησης (MLB)

Μία πρώτη σκέψη είναι να χρησιμοποιήσουμε το σήμα $\bar{r}_t - \bar{\bar{r}}_t$ για τον δυναμικό έλεγχο της αντίστροφης θερμοκρασίας β σε μία προσαρμοστική μηχανή μάθησης μίας κατάστασης για επίλυση του στοχαστικού μη-στατικού προβλήματος MAB. Στο [50] η ιδέα αυτή έχει ήδη χρησιμοποιηθεί σε έναν αλγόριθμο ενισχυτικής μάθησης με κάποιες τροποποιήσεις και γενικότερες παραδοχές. Η διαισθητική προσέγγιση είναι πως η αύξηση των μέσων επιβραβεύσεων ερμηνεύεται ως βελτίωση της επίδοσης και συνεπώς η ανάγκη για εξερεύνηση μειώνεται [50, 27]. Αντιθέτως, πτώσεις της μέσης επιβράβευσης μπορεί να ερμηνευθούν ως δείγματα αλλαγής του περιβάλλοντος μέσω των στατιστικών που το περιγράφουν, και συνεπώς υπάρχει η ανάγκη για αύξηση των εξερευνητικών δράσεων και επαναπροσδιορισμού της γνώσης. Παρ΄ όλα αυτά

η τρέχουσα μέση επιβράβευση $\bar{r}_t$ δεν μπορεί να καθιστά απόλυτη ένδειξη της επίδοσης (π.χ μπορεί να υπάρχουν πολύ καλύτερες επιλογές), και για τον λόγο αυτό απαιτείται μία αναφορά, όπως η τρέχουσα μέση επίδοση 2ης τάξης ή μακρέως χρόνου.

Ως παράδειγμα, σε περιπτώσεις κατά τις οποίες λαμβάνονται μόνο αρνητικές επιβραβεύσεις, η τρέχουσα εκτιμώμενη επιβράβευση θα είναι αρνητική, χωρίς όμως αυτό να συνεπάγεται ότι θα πρέπει να μην υπάρχουν αξιοποιητικές δράσεις καθώς η αρνητική αυτή κατά πρόσημο επίδοση μπορεί να είναι και η βέλτιστη δυνατή. Συνεπώς, ακολουθώντας τις υποθέσεις από το [50], το ρόλο της αναφοράς θα έχει ο τρέχων μέσος 2ης τάξης, ή αλλιώς τρέχων μέσος μακρέως χρόνου $\bar{\bar{r}}_t$, ενώ το ρόλο της τρέχουσας σχετικής επίδοσης θα έχει ο τρέχων μέσος βραχέως-χρόνου $\bar{r}_t$. Όταν $\bar{r}_t > \bar{\bar{r}}_t$, τότε η επίδοση είναι μεγαλύτερη από τη μέση επίδοση και συνεπώς οι δράσεις μπορούν να είναι λιγότερο εξερευνητικές και περισσότερο αξιοποιητικές. Όταν $\bar{r}_t < \bar{\bar{r}}_t$, τότε η επίδοση είναι μικρότερη από την τρέχουσα μέση επίδοση και η ανάγκη για εξερευνητικές δράσεις αυξάνεται.

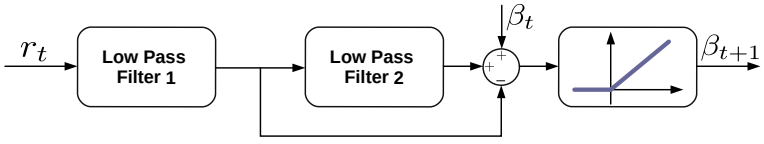
Εάν β_t είναι η μετα-παράμετρος αντίστροφης θερμοκρασίας και $Q_{a,t}$ είναι ο τρέχων μέσος επιβραβεύσεων ανά δράση, τότε η πολιτική λήψης αποφάσεων π θα είναι

$$\pi(a|\beta_t) = \mathbb{P}\left[A_t = a|\beta_t\right] = \frac{\exp\{\beta_t Q_{a,t}\}}{\sum\limits_{a' \in \mathcal{A}} \exp\{\beta_t Q_{a',t}\}} \tag{4.1}$$

Μετά την επιλογή της δράσης A_t και την επιστροφή της επιβράβευσης r_t από το περιβάλλον, η μηχανή μάθησης ανανεώνει τον τρέχοντα μέσο βραχέως χρόνου $\bar{r}_t$ και έπειτα χρησιμοποιεί την τιμή αυτή για να ανανεώσει τον τρέχοντα μέσο μακρέως χρόνου $\bar{\bar{r}}_t$

$$\bar{r}_t = \alpha_m r_t + (1 - \alpha_m)\bar{r}_{t-1} \quad , \quad \bar{\bar{r}}_t = \alpha_\ell \bar{r}_t + (1 - \alpha_\ell)\bar{\bar{r}}_{t-1} \tag{4.2}$$

όπου α_m και α_ℓ είναι παράμετροι που καθορίζουν το ρυθμό μάθησης (learning rates), αντιστρόφως ανάλογοι των σταθερών τ_1 και τ_2 που χρησιμοποιούνται στο [50]. Ο τρέχων μέσος $Q_{A_t,t}$ της δράσης που εκτελέστηκε (στο εξής θα αναφερόμαστε στους τρέχοντες μέσους ως



Σχήμα 4.2: Διάγραμμα μετα-μάθησης της εξερεύνησης στον αλγόριθμο MLB. Το σήμα $\bar{r}_t - \bar{\bar{r}}_t$ καθώς και η προηγούμενη τιμή του β_t χρησιμοποιούνται για την ανανέωση της εξερεύνησης διαμέσου μίας τμηματικά γραμμικής μονάδας (RELU).

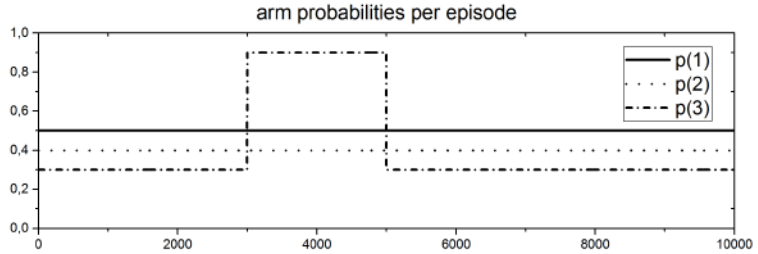
αξίες δράσεων) και η αντίστροφη θερμοκρασία β_t, ανανεώνονται για χρήση τους την επόμενη χρονική στιγμή $t+1$, ως

$$Q_{A_t,t+1} = \alpha_Q r_t + (1-\alpha_Q) Q_{A_t,t} \quad , \quad \beta_{t+1} = \max\{\beta_t + \eta(\bar{r}_t - \bar{\bar{r}}_t), 0\} \tag{4.3}$$

όπου η είναι μία παράμετρος επιλογής ελέγχου της προσαρμοστικότητας και α_Q το βήμα μάθησης. Η πρόταση στο [50] ήταν η χρήση του βήματος μάθησης α_Q ως μιας δυναμικά εξαρτώμενης μετα-παραμέτρου $\alpha_{Q,t}$ απο έναν αντίστοιχο κανόνα με τον κανόνα ανανέωσης του β_t, εμπνευσμένοι από τη βιολογική του συσχέτιση με την ακετυλοχολίνη. Παρά ταύτα, κάτι τέτοιο θα δημιουργήσει την ανάγκη για 3 νέες παραμέτρους (αντίστοιχες των $\alpha_m, \alpha_\ell, \eta$) και η αξιολόγηση της προσαρμοστικότητας καθίσταται εξαιρετικά δύσκολη, λόγω του ότι οι παράμετροι αυτοί χρειάζονται κατάλληλη ρύθμιση και η πολυπλοκότητα του παραμετρικού χώρου αυξάνεται εκθετικά. Η αύξηση αυτή της πολυπλοκότητας του παραμετρικού χώρου αναφέρεται συνήθως και ως η *κατάρα της διαστασιμότητας του Bellman* (Bellman's curse of dimensionality). Στον παραπάνω αλγόριθμο (MLB - meta-learning for bandits) τροποποιούμε δυναμικά μόνο τη μετα-παράμετρο αντίστροφης θερμοκρασίας, κάτι που εμπειρικά βελτιώνει κατά πολύ την επίδοση της απλής soft-max Boltzmann προσέγγισης και ίσως ξεπερνά σε αρκετές περιπτώσεις αυτή των υπόλοιπων σύγχρονων αλγορίθμων.

4.2.1 Αξιολόγηση Αλγορίθμου MLB

Για μία πρώτη αξιολόγηση, η οποία έγινε στο [30] ως μέρος της ανάπτυξης ενός αλγορίθμου ενισχυτικής μάθησης που θα παρουσιαστεί στα επόμενα κεφάλαια, εκτιμήθηκε η επίδοση του απλού MLB αλγορίθμου σε σύγκριση με τους D-UCB, SW-UCB και UCB1 που παρουσιάστηκαν στα κεφάλαια 2, 3. Το κίνητρο πίσω από την ανάπτυξή του ήταν η διερεύνηση των εγγενών προσαρμοστικών χαρακτηριστικών της βιολογικά εμπνευσμένης μετα-μάθησης σε περιβάλλοντα αλληλεπίδρασης ανθρώπου-ρομπότ τα οποία μπορούν να μοντελοποιηθούν

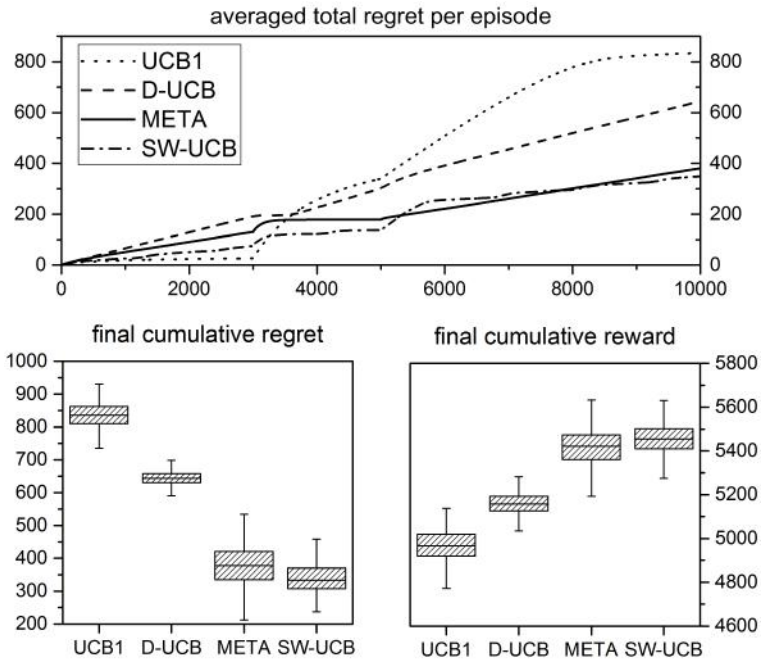


Σχήμα 4.3: Στοχαστικό μη-στατικό περιβάλλον τύπου Bernoulli. Η πιθανότητα επιστροφής μοναδιαίας επιβράβευσης για την τρίτη δράση αλλάζει σε βέλτιστη κατά το χρονικό διάστημα $3000 \leq t < 5000$.

σε μία κατάσταση. Για τα παραπάνω περιβάλλοντα αλληλεπίδρασης ανθρώπου-ρομπότ ίσως σκεφτεί κανείς πως θα ήταν πιο συνετή η μοντελοποίησή τους ως ενός ανταγωνιστικού (μη-στοχαστικού) MAB (εφόσον οι κατανομές των επιβραβεύσεων ενδέχεται να εξαρτώνται από τις παρελθοντικές δράσεις του ρομπότ). Παρόλα αυτά ως αναφορά επιλέχθηκε το ίδιο στοχαστικό και μη-στατικό πρόβλημα που παρουσιάστηκε στο [19] καθώς στα στοχαστικά προβλήματα δεν υπάρχουν συμφραζόμενα (contextual information) που αφορούν τη δυναμική του περιβάλλοντος και συνεπώς η προσαρμοστικότητα του αλγορίθμου εξαρτάται μόνο από την εγγενή προσαρμοστική συμπεριφορά της βιολογικά εμπνευσμένης μετα-μάθησης.

Στο πρόβλημα αυτό υπάρχουν τρεις δράσεις προς επιλογή, η κάθε μία εκ των οποίων επιστρέφει στοχαστικά μία δυαδική επιβράβευση $r_t \in \{0, 1\}$. Το πρόβλημα λοιπόν είναι στοχαστικό, μη-στατικό και τύπου Bernoulli. Εδώ, συμβολίζουμε με $p_t(a)$ την πιθανότητα επιστροφής μοναδιαίας επιβράβευσης $r_t = 1$ κατά την επιλογή της δράσης a, όπου $a \in \{1, 2, 3\}$, ενώ το περιβάλλον ξεκινάει με $p_t(1) = 0.5,\ p_t(2) = 0.4,\ p_t(3) = 0.3$ για $1 \leq t < 3000$, αλλάζει σε $p_t(3) = 0.9$ για $3000 \leq t < 5000$ και επανέρχεται σε $p_t(3) = 0.3$ για $5000 \leq t \leq 10000$, με ορίζοντα $T = 10000$ όπως φαίνεται στο σχήμα 4.3.

Για την υλοποίηση χρησιμοποιήθηκε βήμα μάθησης $\alpha_Q = 0.4$, ενώ η ανανέωση της μεταπαραμέτρου αντίστροφης θερμοκρασίας έγινε σε κάθε βήμα με βάση την εξίσωση 4.3, με την τροποποίηση ότι σε κάθε ανανέωση προστίθεται μικρή σταθερά $\epsilon \simeq 10^{-3}$ ωστέ να διασφαλιστεί η σταδιακή μείωση των εξερευνητικών δράσεων σε περιπτώσεις όπου η απόδοση έχει



Σχήμα 4.4: Συγκριτικά αποτελέσματα αλγορίθμου MLB για το μη-στατικό διακοπτόμενο περιβάλλον του σχήματος 4.3, εκτελώντας 500 επαναλήψεις (ο MLB εδώ έχει το ακρωνύμιο META). Πάνω: Μέση αθροιστική μεταμέλεια ανά χρονική στιγμή. Κάτω αριστερά: Τελική αθροιστική μεταμέλεια. Κάτω δεξιά: Μέση αθροιστική επιβράβευση.

σταθεροποιηθεί λόγω ισότητας των $\bar{r}_t$ και $\bar{\bar{r}}_t$ (κάτι το οποίο μπορεί να συμβεί σε μεγάλα διαστήματα στατικότητας του περιβάλλοντος). Τα υπόλοιπα βήματα μάθησης που χρησιμοποιήθηκαν ήταν $\alpha_m = 1/20,\ \alpha_\ell = 1/300,\ \eta = 75$. Για τη ρύθμιση των υπερπαραμέτων αυτών, χρησιμοποιήθηκε αναζήτηση σε πλέγμα με μεγάλη κλίμακα, παρατηρώντας τις περιοχές του παραμετρικού χώρου όπου η επίδοση ήταν σχετικά καλή και παράλληλα επιδείκνυε ευρωστία, επιλέγοντας την καλύτερη περιοχή προς διερεύνηση, μειώνοντας την κλίμακα και επαναλαμβάνοντας την εξερεύνηση στην περιοχή αυτή. Όλες οι προσομοιώσεις επαναλήφθηκαν για 500 συνεδρίες (sessions) και παρατηρήθηκε η μέση αθροιστική μεταμέλεια των δράσεων καθώς και η μέση αθροιστική επιβράβευση στο σύνολο των συνεδριών.

Όπως φαίνεται από το σχήμα 4.4, ο αλγόριθμος UCB1 σημειώνει καλύτερες επιδόσεις για τις χρονικές στιγμές $1 \leq t < 3000$ (στο εξής η χρονική στιγμή θα αναφέρεται και ως *επεισόδιο*), όπως άλλωστε ήταν και αναμενόμενο από την ανάλυση των προηγούμενων κεφαλαίων, λόγω της στατικής φύσης αυτού του χρονικού διαστήματος. Στα επόμενα επεισόδια όμως αρχίζει και αυξάνεται δραματικά η αθροιστική μεταμέλεια λόγω της υψηλής αδράνειας μάθησης που εκδηλώνει. Ο αλγόριθμος SW-UCB είναι ο δεύτερος καλύτερος κατά τα πρώτα επεισόδια, επιδεικνύοντας ισορροπημένη αναλογία μεταξύ εξερευνητικών και αξοιποιητικών δράσεων, κάτι που είναι εμφανές από τη μικρή κλίση της μέσης αθροιστικής μεταμέλειας στο διάστημα αυτό. Η προσαρμοστική φύση του αλγορίθμου MLB επιδεικνύεται μετά την πρώτη αλλαγή. Η οριζόντια γραμμή για το χρονικό διάστημα από 3500 έως 5000 φανερώνει μάλιστα πως ο αλγόριθμος λειτουργούσε βέλτιστα παρουσιάζοντας μηδενική μεταμέλεια. Στο τέλος του επεισοδίου 5000, ο SW-UCB επιτυγχάνει τη χαμηλότερη αθροιστική μεταμέλεια. Παρ΄ όλα αυτά, η κλίση της γραφικής παράστασης σε αυτό το χρονικό διάστημα είναι σχεδόν η ίδια με αυτή του πρώτου χρονικού διαστήματος. Αυτό σημαίνει πως τα επίπεδα εξερεύνησης γι΄ αυτά τα δύο διαστήματα είναι παρόμοια, παρόλο του μεγαλύτερου χάσματος δράσης που παρατηρείται μεταξύ της βέλτιστης και της δεύτερης καλύτερης δράσης στο δεύτερο διάστημα. Μετά τη δεύτερη αλλαγή, το χάσμα δράσεων είναι και πάλι μικρό, και ο SW-UCB γίνεται και πάλι ο βέλτιστος εκτός από τα τελευταία 1500 επεισόδια στα οποία ο UCB1 ξεπερνάει την αδράνεια μάθησης (ωστόσο ο UCB1 έχει ήδη μεγάλη αθροιστική μεταμέλεια). Τελικά, η συνολική επίδοση του MLB είναι συγκρίσιμη με αυτή του SW-UCB.

Στις συγκεκριμένες προσομοιώσεις οι παράμετροι που χρησιμοποιήθηκαν για τους αλγορίθμους SW-UCB και D-UCB είναι αυτοί που αναφέρονται στο [19] και για τις οποίες εγγυάται η ασυμπτωτική συμπεριφορά της ανάλυσης στο κεφάλαιο 3, ενώ για τον αλγόριθμο MLB οι παράμετροι ρυθμίστηκαν με εκτενή αναζήτηση ώστε να επιτευχθεί η βέλτιστη εμπειρική συμπεριφορά. Παρά ταύτα, στις επόμενες παραγράφους θα παρουσιαστούν εκτενέστερα αριθμητικά αποτελέσματα με συνολική ρύθμιση παραμέτρων για όλους τους αλγορίθμους.

4.3 Υβριδικός Αλγορίθμος Μετα-Μάθησης με Φίλτρα Kalman (MLB-KF)

Στη συνέχεια προτείνεται ένας νέος αλγόριθμος ως μία υβριδική προσέγγιση που ενσωματώνει τους αλγορίθμους KF-MANB και MLB. Έχοντας περιγράψει αναλυτικά τα περισσότερα

τμήματά του, η περιγραφή του μπορεί να γίνει εύκολα με μία απλή αντικατάσταση. Ακολουθώντας τις προτάσεις από τα εξερευνητικά μοντέλα υπολογιστικής νευροεπιστήμης [14, 18] οι αξίες των δράσεων $Q_{a,t}$ της soft-max συνάρτησης στην εξίσωση 4.1 τροποποιούνται ώστε να εμπεριέχουν επιπλέον επιβραβεύσεις εξερεύνησης εξαρτώμενες από την αβεβαιότητα της κάθε δράσης. Για να γίνει πιο σαφές, γίνεται η παρακάτω αντικατάσταση για κάθε δράση a,

$$Q_{a,t} \leftarrow \hat{\mu}_{a,t} + \phi\hat{\sigma}_{a,t}$$

όπου $\hat{\mu}_{a,t}$ και $\hat{\sigma}_{a,t}$ είναι η εκτιμώμενη μέση επιβράβευση και η εκτιμώμενη τυπική της απόκλιση με χρήση των εξισώσεων (3.3), (3.4) του αλγορίθμου KF-MANB του κεφαλαίου 3, ενώ ϕ είναι μία σταθερά επιλογής. Έτσι, η νέα πολιτική π θα είναι

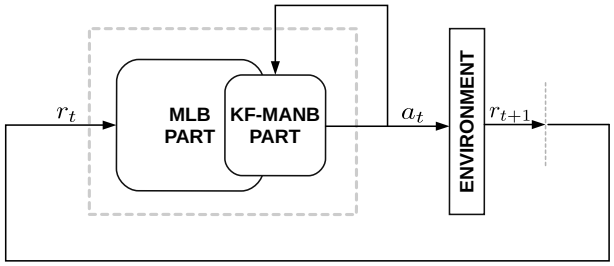
$$\pi(a|\beta_t) = \mathbb{P}\left[A_t = a|\beta_t\right] = \frac{\exp\{\beta_t(\hat{\mu}_{a,t} + \phi\hat{\sigma}_{a,t})\}}{\sum_{a' \in \mathcal{A}} \exp\{\beta_t(\hat{\mu}_{a',t} + \phi\hat{\sigma}_{a',t})\}} \tag{4.4}$$

και η συνολική περιγραφή και τα βήματα του υβριδικού αλγορίθμου MLB-KF φαίνονται στον Αλγόριθμο 1, ενώ ένα διαισθητικό διάγραμμα της διαδικασίας φαίνεται στο σχήμα 4.5.

Η ιδέα της προσθήκης ενός πριμ εξερεύνησης σε κάθε δράση στο εκθετικό της soft-max συνάρτησης δεν είναι καινούρια. Όταν η Gaussian προσέγγιση μιας δράσης έχει υψηλή διασπορά (και συνεπώς αβεβαιότητα), η δράση αυτή θα πρέπει να διερευνηθεί περαιτέρω. Παρόλα αυτά, στον αλγόριθμο KF-MANB, κατά τις στατικές χρονικές περιόδους του περιβάλλοντος, η αθροιστική αβεβαιότητα στις μη επιλεγμένες δράσεις λόγω του θορύβου μετάβασης σ_{tr} επιφέρει άσκοπη εξερεύνηση. Από την άλλη ο απλός αλγόριθμος MLB αναμένεται να επιδεικνύει υψηλή μεταμέλεια στις περιπτώσεις διακοπτόμενων αλλαγών κατά τις οποίες ένα υποβέλτιστο

Algorithm 1 MLB-KF

1: Choose parameters $\alpha_m, \alpha_\ell, \eta, \phi, \sigma_{ob}, \sigma_{tr}$
2: Initialize β_1, and $\hat{\mu}_{a,1}, \hat{\sigma}_{a,1}$ $\forall a \in \mathcal{A}$
3: **for** $t = 1, 2, \ldots, T$ **do**
4: select an action $A_t \in \mathcal{A}$ with Eq. (4.4)
5: observe the immediate reward r_t
6: update $\bar{r}_t, \bar{\bar{r}}_t$, with Eqs. (4.2)
7: for $a = A_t$ update $\hat{\mu}_{a,t+1}, \hat{\sigma}^2_{a,t+1}$ with Eqs. (3.4)
8: for all arms $a \neq A_t$ update $\hat{\mu}_{a,t+1}, \hat{\sigma}^2_{a,t+1}$ with Eqs. (3.3)
9: update β_{t+1} with Eq. (4.3)
10: **end for**



Σχήμα 4.5: Διάγραμμα περιγραφής του αλγορίθμου MLB-KF. Το τμήμα MLB είναι υπεύθυνο για την προσαρμοστικότητα της γενικής εξερεύνησης μέσω του β_t, ενώ το τμήμα KF-MANB εξειδικεύεται στην ειδική εξερεύνηση κάθε δράσης διαμέσου της εκτιμώμενης επίδοσης και της αβεβαιότητάς της.

άκρο μεταβαίνει σε βέλτιστο μετά από μία μεγάλη στατική χρονική περίοδο, καθώς το άκρο αυτό μπορεί με μεγάλη πιθανότητα να μην επιλεγεί για μεγάλο χρονικό διάστημα (κάτι τέτοιο μπορεί επίσης να συμβεί λόγω αριθμητικής υπερχείλισης από τις υψηλές τιμές που λαμβάνουν τα εκθετικά της soft-max). Ο αλγόριθμος MLB-KF δημιουργεί μία ισορροπία μεταξύ των δύο προβλημάτων αυτών, καθώς ενσωματώνει τη γρήγορη προσαρμογή του MLB στις διακοπτόμενες αλλαγές διαφορετικού τύπου από τις παραπάνω, αλλά και την ευρωστία και υψηλή επίδοση του KF-MANB στις υπόλοιπες περιπτώσεις.

Μία σημαντική διαφωνία που θα μπορούσε να έχει κανείς, είναι κατά πόσο είναι λογική και εύλογη η χρήση κανονικών κατανομών σε προβλήματα τύπου Bernoulli. Όμως οι μέσοι των αθροιστικών επιβραβεύσεων θα ακολουθούν κανονική κατανομή λόγω του κεντρικού οριακού θεωρήματος. Έτσι η ερώτηση που τίθεται δεν αφορά απλά το ποια είναι η καλύτερη δράση την επόμενη χρονική στιγμή, αλλά το ποια είναι η καλύτερη δράση γενικά στο μέλλον. Εφόσον αυτή η ερώτηση απαντάται σε κάθε βήμα τότε η δυναμική του περιβάλλοντος μπορεί να ανιχνευθεί επαρκώς, όπως εξάλλου φαίνεται στα εμπειρικά αποτελέσματα που ακολουθούν στην επόμενη παράγραφο.

4.3.1 Αξιολόγηση Υβριδικού Αλγορίθμου MLB-KF

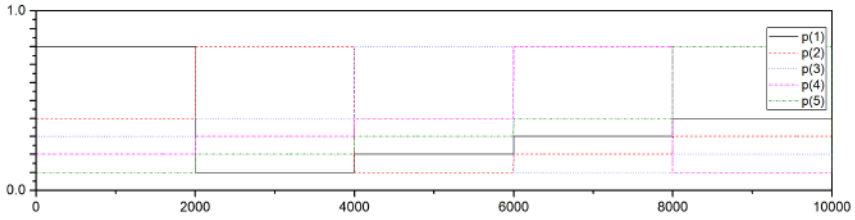
Κατά τη διαδικασία σχεδιασμού ενός τεχνητού περιβάλλοντος με σκοπό την αξιολόγηση μηχανών μάθησης σε προβλήματα που μοντελοποιούνται ως MABs, πρέπει να ληφθούν αρκετά ζητήματα υπόψιν. Στα μη-στατικά περιβάλλοντα με αλλαγές διακοπτόμενου τύπου το σημαντικότερο ζήτημα είναι η επιλογή του ρυθμού με τον οποίο συμβαίνουν οι αλλαγές αυτές. Στο ακόλουθο σύνολο προβλημάτων θα επιλεγεί ένας μικρός ορίζοντας $T = 10000$, όπου το βέλτιστο άκρο[3] θα αλλάζει από 1 έως 10 φορές. Συγκριτικά με άλλα υπάρχοντα προβλήματα που χρησιμοποιούνται για αξιολόγηση, τα προβλήματα που θα χρησιμοποιηθούν μπορεί να θεωρηθούν ως *γρήγορης εναλλαγής*. Ωστόσο θεωρούμε ότι σε πραγματικές εφαρμογές ένα

[3]Υπενθυμίζεται ό,τι οι δράσεις προς επιλογή θα αναφέρονται πολλές φορές και ως *άκρα* (arms) λόγω της αντιστοίχισης των προβλημάτων λήψης αποφάσεων με τη χρήση ενός πλήθους μηχανών επιστροφής στοχαστικής επιβράβευσης, όμοιων με τις μηχανές που χρησιμοποιούνται στα καζίνο (κουλοχέρηδες). Έτσι θεωρούμε ό,τι η κάθε δράση αντιστοιχεί με την επιλογή ενός άκρου κάποιας εκ των μηχανών αυτών.

επεισόδιο μπορεί αντιστοιχηθεί σε μία ημέρα, συνεπώς η ανάγκη για ανάπτυξη προσαρμοστικών αλγορίθμων που επιτυγχάνουν καλή επίδοση σε περιβάλλοντα με υψηλές συχνότητες διακοπτόμενων αλλαγών είναι μεγάλη.

Ένα άλλο σημαντικό ζήτημα σχετίζεται με τον τύπο της διακοπτόμενης αλλαγής του περιβάλλοντος, η οποία μπορεί να είναι *καθολική* (κατά την οποία οι πιθανότητες $p_{t_c}(a)$ όλων των άκρων $a \in \mathcal{A}$ αλλάζουν τη χρονική στιγμή t_c της αλλαγής), ή *ανά άκρο* (κατά την οποία αλλάζει η πιθανότητα $p_{t_c}(a)$ μόνο ενός άκρου). Άλλη μία σημαντική παράμετρος είναι το ελαχιστο χάσμα δράσης Δ_g (αντίστοιχο του d που χρησιμοποιεί ο ϵ-greedy), το οποίο μπορεί να συσχετισθεί με τη *διακριτική ικανότητα* που μπορεί να επιτύχει ένας αλγόριθμος μεταξύ δύο επιλογών. Κάποιες επιπλέον περιπτώσεις προς διερεύνηση είναι αυτές κατά τις οποίες το βέλτιστο άκρο αλλάζει σε υποβέλτιστο μετά από μεγάλη χρονική διάρκεια στατικότητας του περιβάλλοντος, όπως επίσης και περιπτώσεις όπου το χειρότερο εκ των άκρων αλλάζει σε βέλτιστο, ενώ παράλληλα οι στατιστικές των υπόλοιπων άκρων παραμένουν αμετάβλητες. Συνήθως, στις περισσότερες προσομοιώσεις που παρουσιάζονται στη βιβλιογραφία, οι υπερπαράμετροι των αλγορίθμων ρυθμίζονται ώστε να επιτυγχάνεται η βέλτιστη επίδοση, αξιολογώντας όμως πολλές φορές την επίδοση αυτή στο ίδιο σύνολο προβλημάτων που χρησιμοποιήθηκε για τη ρύθμιση. Ενώ κάτι τέτοιο μπορεί πράγματι να δίνει κάποια στοιχεία για την επίδοση του κάθε αλγορίθμου, στη γενικότερη περίπτωση γίνεται *υπερπροσαρμογή* (overfitting) των παραμέτρων στο συγκεκριμένο πρόβλημα και χάνεται η ικανότητα για γενίκευση και επίτευξη αντίστοιχα καλής επίδοσης σε άλλα προβλήματα. Σε αυτή την εργασία προχωρούμε σε ρύθμιση των υπερπαραμέτρων όλων των αλγορίθμων σε ένα συγκεκριμένο στοχαστικό μη-στατικό τεχνητό περιβάλλον 5 άκρων, υπολογίζοντας την επίδοση για το συγκεκριμένο πρόβλημα και ελέγχοντας την ικανότητα γενίκευσης σε διαφορετικά περιβάλλοντα αλλάζοντας αρκετά από τα χαρακτηριστικά του αρχικού, και παρατηρώντας την ευρωστία που επιτυγχάνεται καθώς και την επίδοσης βάσει της αθροιστικής μεταμέλειας.

Όλα τα περιβάλλοντα θα είναι τύπου Bernoulli. Στο σύνολο προβλημάτων τύπου 1, διερευνούμε την επίδοση των αλγορίθμων αλλάζοντας τις πιθανότητες επιστροφής επιβράβευσης κάθε άκρου καθολικά και βάσει τυχαίου περιπάτου, αλλάζοντας την μέση συχνότητα των αλλα-

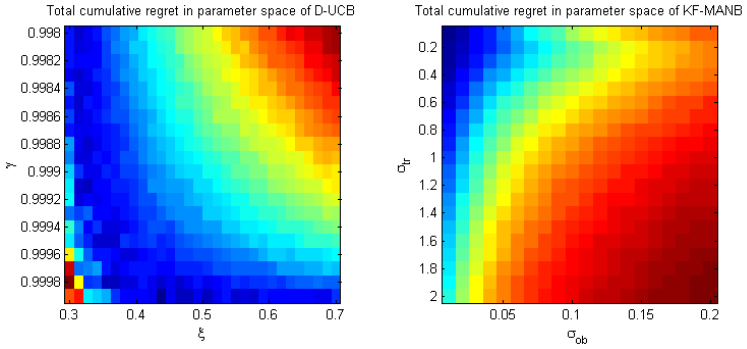


Σχήμα 4.6: Δυναμικό διακοπτόμενο περιβάλλον που χρησιμοποιήθηκε για τη ρύθμιση των παραμέτρων όλων των αλγορίθμων προς διερεύνηση. Οι πιθανότητες επιστροφής μοναδιαίας επιβράβευσης αλλάζουν κυκλικά κάθε 2000 χρονικές στιγμές, ενώ το ελάχιστο χάσμα δράσης παραμένει σταθερό.

γών με στοχαστικό τρόπο. Στα προβλήματα τύπου 2 τροποποιούμε το ελάχιστο χάσμα δράσης Δ_g καθώς επίσης και τον συνολικό αριθμό των καθολικών αλλαγών, με ντετερμινιστικό τρόπο σε αυτή την περίπτωση, ισομοιράζοντας το χρονικό εύρος ζωής του αλγορίθμου σε τμηματικά στατικά διαστήματα. Στο σύνολο προβλημάτων τύπου 3 διερευνούμε την επίδοση σε αλλαγές στις οποίες ένα βέλτιστο άκρο γίνεται υποβέλτιστο, αλλαγές στις οποίες το χειρότερο άκρο γίνεται βέλτιστο, καθώς και συνδυασμούς αυτών.

Α. Ρύθμιση παραμέτρων

Η ρύθμιση παραμέτρων μπορεί να είναι από τις πιο χρονοβόρες διαδικασίες πριν την αξιολόγηση των αλγορίθμων. Όπως αναφέρθηκε κατά την πρώτη αξιολόγηση του MLB, για τους αλγορίθμους D-UCB και SW-UCB προτείνεται η επιλογή τους σε μία κλειστή αναλυτική μορφή στο [19], με την οποία εγγυάται το άνω φράγμα της μεταμέλειας που αναλύθηκε στο κεφάλαιο 3. Οι παράμετροι αυτοί επιλέχθηκαν στο περιβάλλον του σχήματος 4.3 και η συγκριτική τους επίδοση με τον απλό αλγόριθμο βιολογικά εμπνευσμένης μετα-μάθησης αναλύθηκε στο σχήμα 4.4, ωστόσο δεν εγγυώνται τη βέλτιστη εμπειρική επίδοση παρά μόνο την ύπαρξη του άνω φράγματος αυτού. Στα επόμενα σύνολα προβλημάτων, οι παράμετροι που θα χρησιμοποιηθούν θα είναι αυτές για τις οποίες παρατηρείται η καλύτερη εμπειρική μέση επίδοση έτσι ώστε να προβούμε σε ισότιμη μεταχείριση και δίκαιη αξιολόγηση όλων των αλγορίθμων.



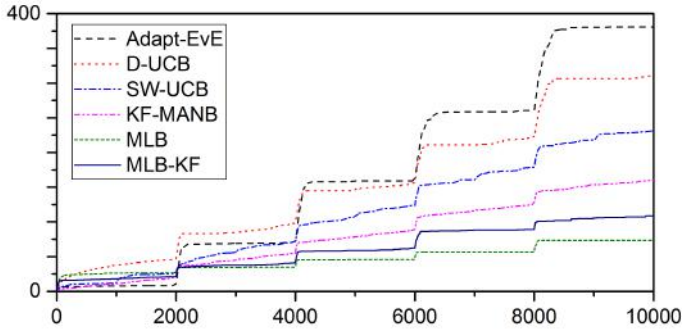
Σχήμα 4.7: Επίδοση αλγορίθμων εντός του παραμετρικού τους χώρου. Αριστερά: Τελική αθροιστική μεταμέλεια που επιτυγχάνει ο D-UCB για διαφορετικές τιμές των παραμέτρων του. Δεξιά: Αντίστοιχη απεικόνιση για τον αλγόριθμο KF-MANB.

Το περιβάλλον στο οποίο θα γίνει η ρύθμιση των παραμέτρων μπορεί να θεωρηθεί ως μία μέση περίπτωση του συνόλου προβλημάτων που μας ενδιαφέρει και στα οποία θα γίνει η τελική αξιολόγηση, αν και δεν υπάρχει κάποια μετρική για μία καλύτερη σαφήνεια του μέσου αυτού. Οι διαθέσιμες δράσεις είναι 5, με αρχικές πιθανότητες επιστροφής μοναδιαίας επιβράβευσης 0.8, 0.4, 0.3, 0.2, 0.1 για την κάθε μία αντίστοιχα, και κυκλική εναλλαγή των πιθανοτήτων αυτών κάθε 2000 χρονικές στιγμές, όπως φαίνεται στο σχήμα 4.6.

Συνολικά επαναλήφθηκαν 200 συνεδρίες κατά τις οποίες υπολογίστηκε η μέση αθροιστική μεταμέλεια στο τέλος του ορίζοντα δοκιμάζοντας διαφορετικές τιμές παραμέτρων για κάθε αλγόριθμο. Οι τιμές των παραμέτρων που δοκιμάστηκαν ήταν αρχικά ορισμένες πάνω σε ένα

αραιό πλέγμα του παραμετρικού χώρου, ενώ στη συνέχεια εξερευνήθηκαν πιο λεπτομερώς οι περιοχές στις οποίες η επίδοση ήταν ικανοποιητική (με την ίδια μέθοδο που αναφέρθηκε κατά την περίπτωση αξιολόγησης του αλγορίθμου MLB).

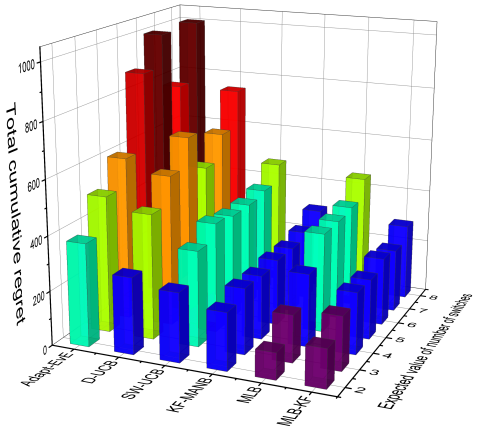
Για τον αλγόριθμο Adapt-EvE η στρατηγική ήταν διαφορετική λόγω της ανάγκης για ρύθμιση του ανιχνευτή διακοπτόμενων αλλαγών (change-point detector). Εάν $x(t)$ είναι μία αρχικοποιημένη ακολουθία μηδενικών τέτοια ώστε $x(t) = 0, \forall t \in [1, 10000]$, τότε για κάθε διαφορετικό ζεύγος των παραμέτρων ευαισθησίας δ, λ του ανιχνευτή αλλαγών Page-Hinkley προσθέταμε μία μονάδα $x(t_c) \leftarrow x(t_c) + 1$ σε κάθε χρονική στιγμή t_c για την οποία ο ανιχνευτής επέστρεφε θετικό αποτέλεσμα. Η διαδικασία αυτή επαναλήφθηκε για 10 συνεδρίες και το τελικό δυαδικό σήμα ομαλοποιήθηκε με ένα παράθυρο τύπου Hanning, οπού $\tilde{x}(t)$ το τελικό σήμα. Το δυαδικό σήμα $y(t)$, το οποίο είναι μηδενικό για κάθε χρονική στιγμή που δεν υπάρχει πραγματική αλλαγή, και μονάδα στις χρονικές στιγμές που πραγματοποιούνται οι αλλαγές, ομαλοποιήθκε επίσης με ένα παράθυρο Hanning και έστω $\tilde{y}(t)$ το αποτέλεσμα. Στη συνέχεια υπολογίστηκε ο συντελεστής συσχέτισης μεταξύ των $\tilde{x}(t)$ και $\tilde{y}(t)$ και επαναλαμβάνοντας τη διαδικασία για όλα τα ζεύγη δ, λ εντός του παραμετρικού χώρου, επιλέχθηκε τελικά το ζεύγος για ο οποίο ο ανιχνευτής επιδείκνυε την καλύτερη δυνατή αναγνώριση αλλαγών. Ώς βασικός αλγόριθμος λήψης απόφασης χρησιμοποιήθηκε ο UCBT, ενώ το εύρος του χρονικού παραθύρου t_m του meta-bandit επιλέχθηκε επισης εμπειρικά. Στο σχήμα 4.7 φαίνεται ενδεικτικά η τελική αθροιστική μεταμέλεια στον παραμετρικό χώρο των αλγορίθμων D-UCB και KF-MANB.



Σχήμα 4.8: Επίδοση των αλγορίθμων μετά τη ρύθμιση των παραμέτρων τους, στο ίδιο πρόβλημα που χρησιμοποιήθηκε για τη ρύθμιση αυτή.

Οι παράμετροι που επιλέχθηκαν τελικώς για τον D-UCB ήταν $\xi = 0.22$ και $\gamma = 0.9999$, για τον SW-UCB $\xi = 0.3$ και $\tau = 998$, για τον Adapt-EvE $\delta = 0.13,\ \lambda = 40,\ t_m = 50$ και χρησιμοποιήθηκε ο UCB-Tuned. Για τον KF-MANB $\sigma_{ob} = 0.2,\ \sigma_{tr} = 0.01$, αρχικοποιώντας όλους του μέσους και τις διασπορές στο 0.5. Για τον MLB $\alpha_Q = 0.14,\ \alpha_m = 1/15,\ \alpha_\ell = 1/350,\ \eta = 0.44$. Για τον MLB-KF χρησιμοποιήθηκαν οι αντίστοιχες παράμετροι από τους MLB και KF-MANB ενώ ρυθμίστηκε μόνο το $\phi = 1.5$. Οι τιμές των παραμέτρων αυτές διατηρήθηκαν σταθερές για όλο τα υπόλοιπα σύνολα προβλημάτων. Από τα αποτελέσματα της μέσης αθροιστικής μεταμέλειας ανά χρονική στιγμή για το σύνολο των 200 συνεδριών, όπως φαίνονται στο σχήμα 4.8, ο αλγόριθμος MLB πετυχαίνει την καλύτερη επίδοση, επιδει-

κνύοντας αξιοποιητική συμπεριφορά όταν είναι απαραίτητο, όπως φαίνεται από την τμηματικά οριζόντια γραμμή, αλλά και εξερευνητική συμπεριφορά στις αλλαγές του περιβάλλοντος, όπως φαίνεται από τη γρήγορη προσαρμογή στις αλλαγές. Ο KF-MANB είναι ο τρίτος καλύτερος, ενώ η επίδοση του MLB-KF είναι μεταξύ των δύο. Ο SW-UCB φαίνεται να πραγματοποιεί αρκετές εξερευνητικές δράσεις, ενώ ο D-UCB εμφανίζει αρκετή αδράνεια μάθησης και αργεί να προσαρμοστεί στις αλλαγές (αν και μετά την προσαρμογή του έχει καλύτερη επίδοση). Ο Adapt-EvE επιδεικνύει τη μεγαλύτερη μεταμέλεια στις αλλαγές, κυρίως λόγω της χρονικής περιόδου λειτουργίας του meta-bandit η οποία είναι απαραίτητη. Μετά από την προσαρμογή του ωστόσο επιτυγχάνει πολύ μικρή μεταμέλεια δράσεων, το οποίο είναι και ο βασικός λόγος για τον οποίο έχει δείξει πολύ καλά αποτελέσματα σε περιβάλλοντα με μεγαλύτερα διαστήματα στατικότητας από αυτά που μελετάμε εδώ (όπως στο σύνολο προβλημάτων Pascal EvE). Αναφορικά με τις διασπορές (οι οποίες δεν φαίνονται στο σχήμα), ο D-UCB και ο Adapt-EvE είχαν μεγαλύτερη διασπορά από τους υπόλοιπους, ενώ ο KF-MANB είχε τη μικρότερη και ο MLB-KF τη δεύτερη μικρότερη. Ο MLB είχε πολλά έκτοπα αποτελέσματα (outliers), ενώ ο MLB-KF όχι. Συνολικά σε αυτή την πρώτη φάση ο MLB-KF επέδειξε στοιχεία ισορροπίας μεταξύ της ευρωστίας του KF-MANB και της καλής μέσης επίδοσης του MLB.



Σχήμα 4.9: Επίδοση των αλγορίθμων για το 1ο σύνολο προβλημάτων. Οι μπάρες αντιπροσωπεύουν τη συνολική αθροιστική μεταμέλεια στο τέλος του ορίζοντα υπολογισμένη κατά μέση τιμή για όλες τις υπερ-συνεδρίες. Ο αλγόριθμος MLB-KF φαίνεται να συνδυάζει τα καλά χαρακτηριστικά αμφότερων των KF-MANB και MLB

Β. Σύνολο προβλημάτων τύπου 1

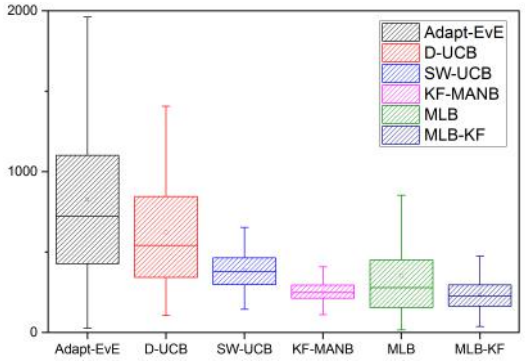
Σε αυτό το σύνολο προβλημάτων, θεωρούμε ό,τι σε κάθε χρονική στιγμή t μπορεί να συμβεί μία καθολική διακοπτόμενη αλλαγή cp του περιβάλλοντος με κάποια πιθανότητα h. Όταν συμβεί αλλαγή, όλες οι νέες πιθανότητες επιβράβευσης των άκρων δειγματοληπτούνται από την ομοιόμορφη κατανομή στο [0,1]. Για κάθε υποπρόβλημα που δημιουργείται από αυτή τη διαδικασία, αξιολογούμε κάθε αλγόριθμο για 20 συνεδρίες δημιουργούμε ένα νέο υποπρόβλημα για την ίδια τιμή h και επαναλαμβάνουμε όλη τη διαδικασία για συνολικά 100 υπερ-συνεδρίες,

υπολογίζοντας την τελική αθροιστική μεταμέλεια κάθε υπερ-συνεδρίας και τη μέση αθροιστική μεταμέλεια από τις 100 υπερ-συνεδρίες αυτές. Στη συνέχεια επαναλαμβάνουμε για διαφορετικές τιμές πιθανότητας, οι οποίες συνολικά επιλέχθηκαν έτσι ώστε $h = n/10000$ με $n = \{2, 3, \ldots, 8\}$, δηλαδή για 2 έως και 8 εκτιμώμενες καθολικές διακοπτόμενες αλλαγές του περιβάλλοντος.

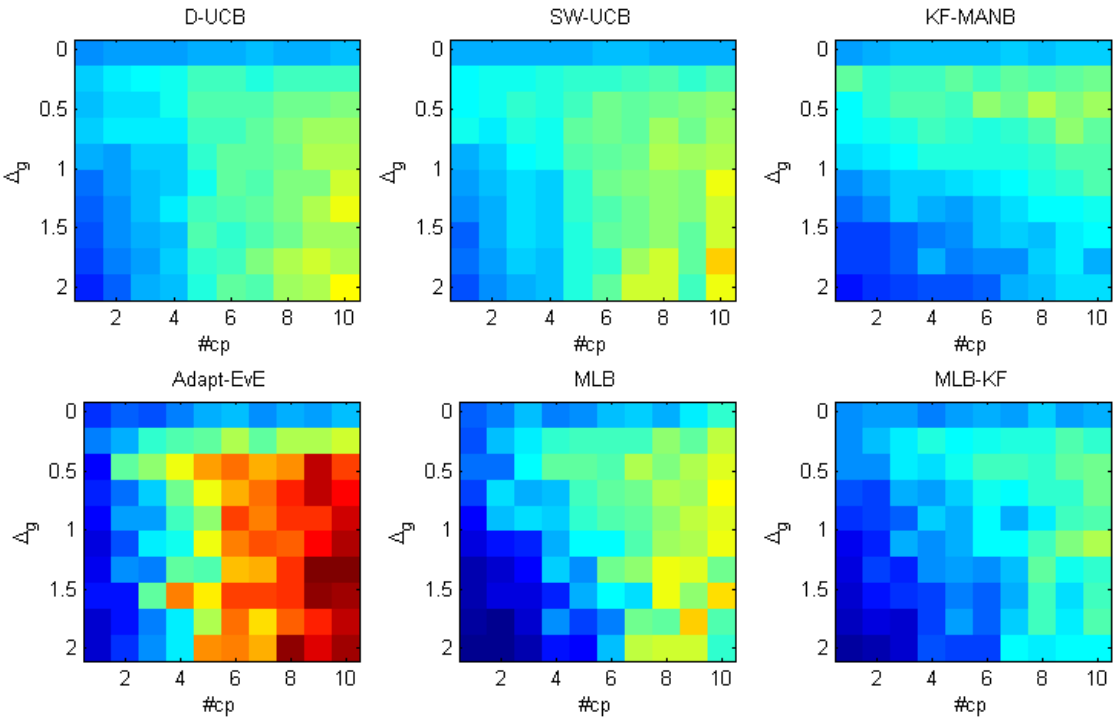
Από τα αποτελέσματα που φαίνονται στο σχήμα 4.9 φαίνεται ότι ο MLB είχε την καλύτερη επίδοση για μικρές τιμές του h, ενώ ο MLB-KF ήταν ο δεύτερος καλύτερος. Όταν ο ρυθμός διακοπών αυξήθηκε, ο MLB επέδειξε μεγάλη διασπορά στην απόδοσή του (δεν φαίνεται στο σχήμα) και η μεταμέλεια αυξήθηκε. Ο KF-MANB ήταν γενικά εύρωστος σε όλες τις περιπτώσεις, πετυχαίνοντας κατά μέσο όρο καλή επίδοση για όλες τις τιμές του h. Ο MLB-KF ενσωματώνει την καλή συμπεριφορά του MLB για χαμηλές συχνότητες αλλαγών, όπως επίσης την ευρωστία και την επίδοση του KF-MANB στις υπόλοιπες περιπτώσεις, βελτιώνοντας ελαφρώς την επίδοση. Ο Adapt-EvE είχε πολύ υψηλή μεταμέλεια (καθώς και μεγάλη διασπορά) ενώ η απόδοσή του μειώθηκε δραματικά με την αύξηση του h. Κάτι παρόμοιο συνέβη και με τον D-UCB. Ο SW-UCB είχε συνολικά καλή και σταθερή επίδοση, ωστόσο συνολικά είναι εμφανές ό,τι ο MLB-KF είχε την καλύτερη μέση επίδοση, κάτι που φαίνεται και στο σχήμα 4.10, στο οποίο απεικονίζονται οι κατανομές της τελικής αθροιστικής μεταμέλειας κάθε αλγορίθμου για όλες τις τιμές του h.

Γ. Σύνολο προβλημάτων τύπου 2

Σε αυτό το σύνολο προβλημάτων τροποποιούμε το ελάχιστο χάσμα δράσης Δ_g, δηλαδή τη διαφορά μεταξύ των πιθανοτήτων επιστροφής μοναδιαίας επιβράβευσης μεταξύ της βέλτιστης και της δεύτερης κατά σειρά βέλτιστης δράσης, ενώ παράλληλα εναλλάσσουμε κυκλικά τις πιθανότητες όλων των άκρων. Το βέλτιστο άκρο θα έχει πάντα πιθανότητα 0.8, το δεύτερο βέλτιστο 0.8-Δ_g, το τρίτο 0.8-3Δ_g και ούτω καθεξής. Εάν $\#cp$ είναι το πλήθος των αλλαγών του περιβάλλοντος στο εύρος του ορίζοντα, στα πειράματα αυτά ξεκινήσαμε με $\#cp = 1$, δηλαδή μία αλλαγή η οποία εκδηλώθηκε τη χρονική στιγμή $t = T/2$ κατά την οποία η πιθανότητα του βέλτιστου άκρου έπεσε από 0.8 σε 0.8-Δ_g, του δεύτερου βέλτιστου από 0.8-Δ_g σε 0.8-2Δ_g



Σχήμα 4.10: Κατανομή αθροιστικής μεταμέλειας στα προβλήματα τύπου 1, για όλες τις διαφορετικές τιμές h που δοκιμάστηκαν.



Σχήμα 4.11: Επίδοση των αλγορίθμων για το 1ο σύνολο προβλημάτων MAB. Πάνω Αριστερά: Τελική αθροιστική μεταμέλεια για τον D-UCB, για επιλογή διαφορετικού ελάχιστου χάσματος δράσης Δ_g και αριθμού αλλαγών $\#cp$. Πάνω Μέση: SW-UCB, Πάνω Δεξιά: KF-MANB, Κάτω Αριστερά: Adapt-EvE, Κάτω Μέση: MLB, Κάτω Δεξιά: D-UCB. Το κόκκινο χρώμα αντιπροσωπεύει υψηλή μεταμέλεια ενώ το μπλε χαμηλή. Έχουν χρησιμοποιηθεί οι ίδιες αναφορές για τη συγκριτική απεικόνιση.

και ούτω καθεξής, με τη διαφορά ό,τι η πιθανότητα του χειρότερου άκρου άλλαξε από 0.8-4Δ_g σε 0.8, δηλαδή στο νέο βέλτιστο άκρο. Στα πειράματα αυτά δοκιμάσαμε διαφορετικές τιμές για το Δ_g από 0.02 έως 0.2 με βήμα 0.02. Για τον κάθε αλγόριθμο έγινε προσομοίωση 200 συνεδριών και υπολογίστηκε η μέση τελική αθροιστική μεταμέλεια. Έπειτα αυξήσαμε το $\#cp$ κατά ένα, καθορίσαμε τις χρονικές στιγμές των αλλαγών έτσι ώστε να είναι ισομοιρασμένες στο εύρος του ορίζοντα και επαναλάβαμε την παραπάνω διαδικασία. Στο σχήμα 4.11 φαίνεται η απεικόνιση της επίδοσης κάθε αλγορίθμου για διαφορετικές τιμές του Δ_g και του $\#cp$.

Από τα αποτελέσματα αυτά φαίνεται πως ο αλγόριθμος MLB-KF και πάλι συνδυάζει τα θετικά των MLB και KF-MANB. Η κάτω αριστερά μπλέ περιοχή προβλημάτων που αφορά την καλή επίδοση του MLB, κληρονομήθηκε από τον MLB-KF αλγόριθμο. Η ικανοποιητική επίδοση του KF-MANB στη δεξιά περιοχή προβλημάτων κληρονομήθηκε επίσης από τον MLB-KF. Ο D-UCB και ο SW-UCB επέδειξαν παρόμοια συμπεριφορά, ενώ ο Adapt-EvE επέδειξε πολύ καλή επίδοση για μικρό πλήθος αλλαγών, αλλά δραματική μείωση της επίδοσης αυτής στις υψηλότερες συχνότητες. Για τη συνολική εκτίμηση της επίδοσης παρατηρήσαμε και τις κατανομές τις τελικής αθροιστικής μεταμέλειας συνολικά για όλα τα Δ_g και όλα τα $\#cp$, και

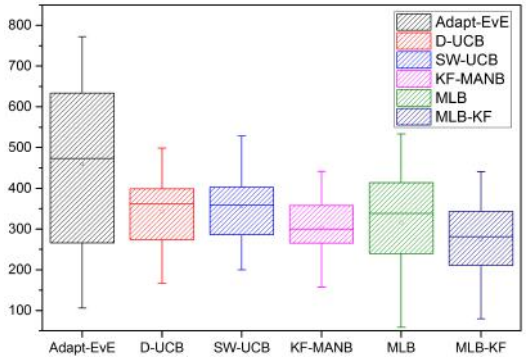
τα αποτελέσματα φαίνονται στο σχήμα 4.12. Ο MLB-KF επιτυγχάνει τη χαμηλότερη μέση μεταμέλεια με ικανοποιητική διασπορά, καλύτερη από αυτή του MLB αλλά λίγο χειρότερη από αυτή του KF-MANB. Τέλος, ο D-UCB είχε γενικά καλύτερη επίδοση από αυτή που επέδειξε στο πρώτο σύνολο προβλημάτων.

Δ. Σύνολο προβλημάτων τύπου 3

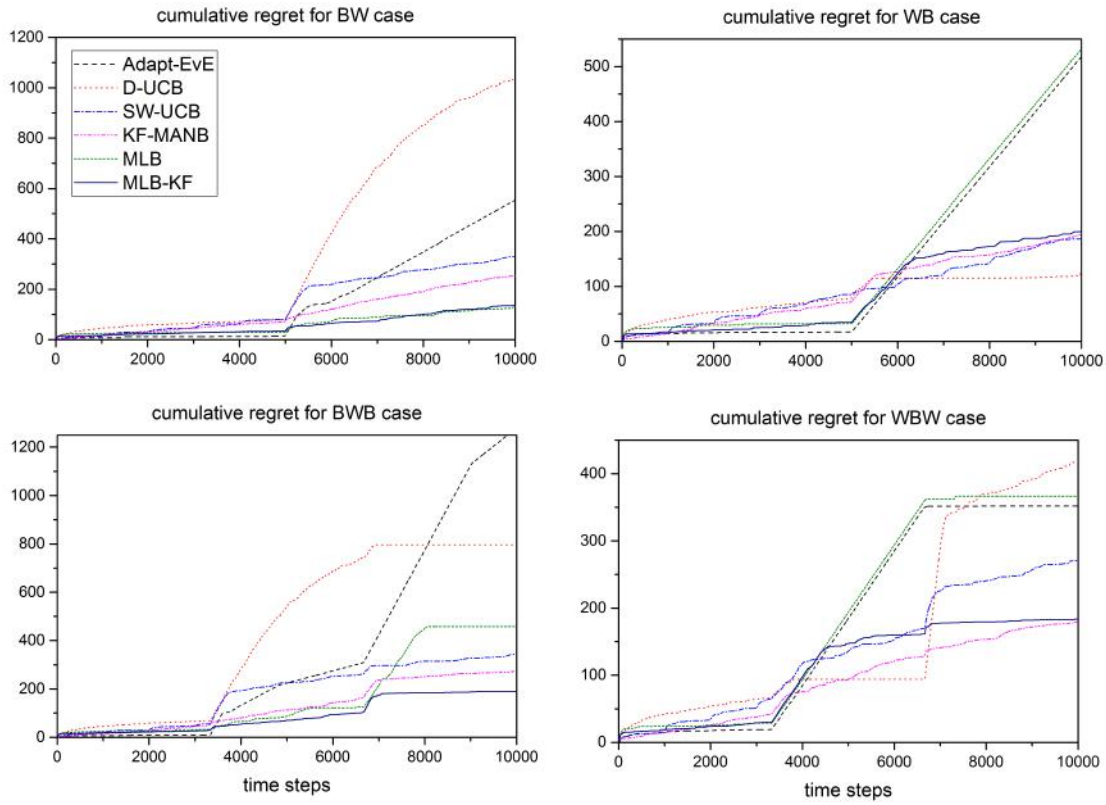
Στο τελευταίο σύνολο προβλημάτων ο σκοπός ήταν η παρατήρηση της επίδοσης σε διαφορετικού τύπου σενάρια αλλαγών. Στα σενάρια BW (Best to Worst), η βέλτιστη δράση αλλάζει σε αυτή που επιφέρει τα χαμηλότερα κέρδη ενώ στα σενάρια WB (Worst to Best) η χειρότερη δράση γίνεται βέλτιστη. Με το ίδιο σκεπτικό μπορούμε να παρατηρήσουμε και τα περιβάλλοντα τύπου BWB ή/και WBW. Οι αρχικές τιμές πιθανοτήτων των άκρων ήταν 0.8, 0.5, 0.4, 0.3, 0.2. Στα σενάρια BW, η πιθανότητα 0.8 του βέλτιστου άκρου αλλάζει απότομα σε 0.1 τη χρονική στιγμή $t = T/2$. Στα σενάρια WB το χειρότερο άκρο με πιθανότητα 0.2 αλλάζει σε 0.9 και γίνεται βέλτιστο την χρονική στιγμή $t = T/2$. Στα BWB το βέλτιστο άκρο αλλάζει από 0.8 σε 0.1 τη χρονική στιγμή $t = T/3$ και επανέρχεται σε 0.8 τη χρονική στιγμή $t = 2T/3$. Στα σενάρια WBW το χειρότερο άκρο αλλάζει από 0.1 σε 0.9 τη χρονική στιγμή $t = T/3$ και επανέρχεται σε 0.1 τη χρονική στιγμή $t = 2T/3$. Συνολικά έγινε προσομοίωση για 200 υπερ-συνεδρίες για κάθε ένα από τα 4 παραπάνω σενάρια και υπολογίστηκε η αθροιστική μεταμέλεια για κάθε χρονική στιγμή, όπως φαίνεται στο σχήμα 4.13

Στα σενάρια BW, οι MLB και MLB-KF επιδεικνύουν την καλύτερη επίδοση. Αυτό συμβαίνει κυρίως λόγω του ότι η απότομη πτώση των άμεσων επιβραβεύσεων οδηγεί σε μείωση της τρέχουσας μέσης επιβράβευσης βραχέως χρόνου $\bar{r}_t$, ενώ ο αντίστοιχος μέσος μακρέως χρόνου $\bar{\bar{r}}_t$ έχει μεγαλύτερη αδράνεια και η τιμή του θα παραμείνει υψηλή για περισσότερα επεισόδια. Συνεπώς η διαφορά $\bar{r}_t - \bar{\bar{r}}_t$ γίνεται μικρότερη του μηδενός και το β_t μειώνεται, αυξάνοντας τα επίπεδα εξερευνητικών δράσεων. Ο KF-MANB ήταν ο δεύτερος καλύτερος, ενώ ο D-UCB επιδεικνύει και πάλι υψηλή αδράνεια μάθησης.

Στα σενάρια WB η μεταμέλεια του MLB είναι πολύ υψηλή. Απο την αντίστοιχη γραφική παράσταση φαίνεται ότι η πολιτική λήψης απόφασης δεν προσαρμόστηκε καθόλου στην αλλαγή.



Σχήμα 4.12: Κατανομή αθροιστικής μεταμέλειας στα προβλήματα τύπου 2, για όλα διαφορετικά ελάχιστα χάσματα δράσης Δ_g και πλήθος αλλαγών του περιβάλλοντος $\#cp$.



Σχήμα 4.13: Επιδόσεις αλγορίθμων στα προβλήματα τύπου 3. Πάνω αριστερά: αθροιστική μεταμέλεια για αλλαγές τύπου BW. Πάνω δεξιά: αθροιστική μεταμέλεια για αλλαγές τύπου WB. Κάτω αριστερά: αθροιστική μεταμέλεια για αλλαγές τύπου BWB. Κάτω δεξιά: αθροιστική μεταμέλεια για αλλαγές τύπου WBW

Το ίδιο φαίνεται ότι συνέβη και στην περίπτωση του Adapt-EvE, ενώ ο D-UCB πέτυχε την καλύτερη επίδοση. Ο SW-UCB και ο KF-MANB είχαν παρόμοιες αποδόσεις. Από τα αποτελέσματα φαίνεται ό,τι ο MLB-KF είχε αρχικά παρόμοια πολιτική με αυτή του MLB, όμως όταν ο MLB άρχισε να έχει χαμηλή επίδοση, ο MLB-KF προσάρμοσε την πολιτική του σε παρόμοια με αυτή του KF-MANB. Στο σενάριο αυτό επιδεικνύεται η υβριδική φύση του MLB-KF με τον καλύτερο δυνατό τρόπο.

Στα σενάρια BWB φαίνεται να συνδυάζονται οι αποδόσεις των δύο κατηγοριών. Ο MLB-KF επιδεικνύει την καλύτερη δυνατή επίδοση, ενώ ο KF-MANB έρχεται ο δεύτερος και ακολουθεί ο SW-UCB. Ο MLB προσαρμόστηκε γρήγορα στην πρώτη αλλαγή (εφόσον η πρώτη αλλαγή είναι τύπου BW), όμως η μεταμέλεια αυξήθηκε ραγδαία μετά τη δεύτερη αλλαγή (εφόσον η δεύτερη αλλαγή είναι τύπου WB). Τέλος, στο σενάριο WBW ο MLB-KF επιδεικνύει παρόμοια συμπεριφορά με αυτή του KF-MANB, ενώ οι δυο τους πετυχαίνουν την καλύτερη επίδοση για τους λόγους που αναφέρθηκαν και στα προηγούμενα σενάρια.

4.4 Συνολική Αξιολόγηση και Συζήτηση

Ο έλεγχος της επίδοσης ενός αλγορίθμου στο ίδιο πρόβλημα που χρησιμοποιήθηκε για τη ρύθμιση των παραμέτρων του, περιορίζει την αξιολόγηση της απόδοσής του. Στα προηγούμενα σύνολα προβλημάτων, αποφύγαμε αυτού του τύπου τη μεροληψία εκτιμώντας εμπειρικά την ικανότητα γενίκευσης του κάθε αλγορίθμου αλλάζοντας τα πιο σημαντικά χαρακτηριστικά ενός στοχαστικού και μη-στατικού περιβάλλοντος και χρησιμοποιώντας τις παραμέτρους οι οποίες ρυθμίστηκαν σε ένα μέσο πρόβλημα. Παρ΄ όλα αυτά αξίζει να σημειώσουμε ότι όταν η αξιολόγηση έγινε στο ίδιο περιβάλλον με αυτό που χρησιμοποιήθηκε για ρύθμιση, ο απλός αλγόριθμος MLB είχε την καλύτερη επίδοση, ενώ ο υβριδικός MLB-KF τη δεύτερη καλύτερη.

Εάν δεν λάβουμε υπόψιν την επίδοση του υβριδικού αλγορίθμου, η γενική εικόνα είναι ότι οι αλγόριθμοι MLB και KF-MANB είναι οι δύο αποδοτικότεροι. Ο MLB παρουσιάζει χαμηλή μεταμέλεια σε πολλές περιπτώσεις, όμως παρουσιάζει επίσης υψηλή διασπορά και χαμηλή ευρωστία. Από την άλλη, ο KF-MANB επιδεικνύει την πιο εύρωστη συμπεριφορά επιτυγχάνοντας παράλληλα πολύ καλή επίδοση. Η ισορροπημένη επιλογή των καλύτερων χαρακτηριστικών των δύο αλγόριθμων αυτών, επιτυγχάνεται με τον υβριδικό αλγόριθμο που αναπτύχθηκε, τον οποίο και ονομάσαμε MLB-KF.

Στη φύση, οι καθημερινού τύπου εφαρμογές που απαιτούν τη λήψη αποφάσεων υπό αβεβαιότητα, έχουν συνήθως μεγαλύτερο ρυθμό αλλαγών από αυτόν που έχει μελετηθεί στη βιβλιογραφία. Η αξιολόγηση των αλγορίθμων λήψης αποφάσεων σε περιβάλλοντα με διαφορετικούς ρυθμούς διακοπτόμενων αλλαγών, ήταν μία από τις πολλές περιπτώσεις που διερευνήθηκαν. Πιο συγκεκριμένα, υπό την παρουσία καθολικού τύπου αλλαγών στο πρώτο σύνολο προβλημάτων ο απλός αλγόριθμος με εμπνευσμένη βιολογική μετα-μάθηση MLB, επέδειξε την καλύτερη επίδοση για χαμηλούς ρυθμούς αλλαγών, ενώ ο αλγόριθμος KF-MANB επέδειξε καλύτερη επίδοση και ευρωστία σε υψηλούς ρυθμούς. Ο υβριδικός αλγόριθμος MLB-KF κληρονομεί τα καλύτερα χαρακτηριστικά τους και επιτυγχάνει μεταμέλεια παρόμοια με τη βέλτιστη που επιτυγχάνεται από τους δύο σε κάθε χρονική στιγμή.

Κάποιες περιπτώσεις που επίσης έχουν ενδιαφέρον προς διερεύνηση, είναι αυτές στις οποίες οι εκτιμώμενες επιβραβεύσεις από δράση σε δράση έχουν μεγάλη ή μικρή διαφορά. Στις περιπτώσεις που τα χάσματα δράσης είναι μικρά, η στιγμιαία μεταμέλεια κάθε υποβέλτιστης δράσης θα είναι και αυτή μικρή, αλλά η αθροιστική μεταμέλεια θα αυξάνεται γραμμικά με την πάροδο του χρόνου. Για τη διερεύνηση αυτών των περιπτώσεων αναπτύξαμε το δεύτερο σύνολο προβλημάτων, στο οποίο τροποποιούσαμε το ελάχιστο χάσμα δράσης, θεωρώντας ότι το χάσμα αυτό μπορεί να συσχετισθεί με τη διακριτική ικανότητα του κάθε αλγορίθμου. Παράλληλα, τροποποιούσαμε και το ρυθμό των διακοπτόμενων αλλαγών με έναν πιο ντετερμινιστικό τρόπο από αυτόν που χρησιμοποιήθηκε στο πρώτο σύνολο προβλημάτων, ώστε να παρατηρήσουμε τυχόν συσχέτιση των δύο χαρακτηριστικών διαμέσου της επίδοσης. Από τις προσομοιώσεις που έγιναν εξήγαμε το συμπέρασμα ότι ο υβριδικός MLB-KF είχε τη συνολικά καλύτερη συμπεριφορά, επιτυγχάνοντας καλή επίδοση αλλά και καλή ευρωστία.

Για να γίνει η αξιολόγηση της πολιτικής κάθε αλγορίθμου αναφορικά με το πόσο αποδοτικά επαναπροσδιορίζει το κλάσμα εξερεύνησης-αξιοποίησης μετά από αλλαγές του περιβάλλοντος,

αναπτύχθηκε το τρίτο σύνολο προβλημάτων, με περιβάλλοντα στα οποία η βέλτιστη επιλογή γίνεται υποβέλτιστη, περιβάλλοντα όπου μία υποβέλτιστη επιλογή γίνεται βέλτιστη, καθώς και συνδυασμοί αυτών. Ο MLB έδειξε αρκετά καλή προσαρμοστικότητα στην πρώτη περίπτωση και ήταν ο καλύτερος εκ των αλγορίθμων, ενώ ο D-UCB στη δεύτερη. Ωστόσο ο κάθε ένας από αυτούς είχε κακή επίδοση στο περιβάλλον όπου ο δεύτερος ήταν ο βέλτιστος. Ο MLB-KF είχε συνολικά τη μέση καλύτερη επίδοση.

Επίσης, εξετάστηκαν σενάρια στα οποία οι αλλαγές ήταν ανά άκρο και όχι καθολικές, όπως επίσης και σενάρια με ομαλά μη-στατικά περιβάλλοντα. Ο MLB-KF είχε επίσης την καλύτερη μέση επίδοση των υπολοίπων. Αν και οι αντίστοιχες προσομοιώσεις δεν απεικονίζονται εδώ, στα επόμενα κεφάλαια που αφορούν δυναμική μάθηση σε περισσότερες καταστάσεις θα γίνει εμφανής η αποδοτικότητα της βιο-εμπνευσμένης μετα-μάθησης σε προβλήματα και αυτού του τύπου.

Κεφάλαιο 5

Ενισχυτική Μάθηση

Σε αυτό το κεφάλαιο θα περιγραφεί η βασική θεωρία υπό το πλαίσιο των Μαρκοβιανών Διαδικασιών Λήψης Αποφάσεων (MDPs) και την Ενισχυτική Μηχανική Μάθηση [56].

5.1 Μαρκοβιανές Διαδικασίες Λήψης Αποφάσεων

Το κατάλληλο μαθηματικό πλαίσιο περιγραφής ενός περιβάλλοντος ενισχυτικής μάθησης είναι οι Μαρκοβιανές διαδικασίες λήψης αποφάσεων (MDPs). Ωστόσο για να μπορεί να συμβεί κάτι τέτοιο, θα πρέπει το περιβάλλον να είναι πλήρως παρατηρήσιμο, ώστε η κάθε κατάσταση να εμπεριέχει τα απαραίτητα χαρακτηριστικά για την πλήρη περιγραφή και ανάκτηση των δυναμικών μεταβάσεων και αλληλεπιδράσεων που μπορούν να συμβούν. Χωρίς βλάβη της γενικότητας μπορούμε να δηλώσουμε ότι τα περισσότερα προβλήματα ενισχυτικής μάθησης μπορούν να μοντελοποιηθούν ως MDPs. Για παράδειγμα, ο βέλτιστος έλεγχος αφορά MDPs μη πεπερασμένων καταστάσεων (συνεχή MDPs), τα multi-armed bandits που μελετήσαμε στα πρώτα κεφάλαια είναι MDPs μίας κατάστασης, και ακόμα και κάποια προβλήματα μερικής παρατηρησιμότητας μπορούν επίσης να μετατραπούν σε MDPs. Για τη μοντελοποίηση των μεταβάσεων και των αλληλεπιδράσεων με αυτόν τον τρόπο, θα πρέπει οι πιθανές καταστάσεις του περιβάλλοντος να ικανοποιούν τη Μαρκοβιανή ιδιότητα,

Ορισμός 5.2. *Η κατάσταση S_t είναι Markov εάν και μόνο εάν,*

$$\mathbb{P}[S_{t+1}|S_1, \ldots, S_t] = \mathbb{P}[S_{t+1}|S_t]$$

η οποία στην ουσία της δηλώνει την *ανεξαρτησία του μέλλοντος από το παρελθόν, δεδομένου του παρόντος*. Στις περιπτώσεις που η κατάσταση δεν είναι Μαρκοβιανή, ενδέχεται να μπορούμε να κατασκευάσουμε έναν νέο επαυξημένο χώρο κατάστασης ώστε να επιτευχθεί η Μαρκοβιανή ιδιότητα. Για παράδειγμα, αν η πιθανότητα μετάβασης εξαρτάται από τις δύο προηγούμενες καταστάσεις έτσι ώστε $\mathbb{P}[S_{t+1}|S_1, \ldots, S_t] = \mathbb{P}[S_{t+1}|S_t, S_{t-1}]$, τότε αρκεί να θέσουμε ως νέα

κατάσταση την κλίκα $\tilde{S}_t = \{S_t, S_{t-1}\}$. Κάτι τέτοιο είναι πολύ χρήσιμο, αν και όχι πάντα εφικτό καθώς δεν γνωρίζουμε το εύρος της ιστορίας που είναι απαραίτητο για την επίτευξη της εν λόγω ανεξαρτησίας.

Για την μαθηματική περιγραφή των Μαρκοβιανών διαδικασιών λήψης αποφάσεων, είναι σκόπιμο να ξεκινήσουμε την περιγραφή από τα χαμηλότερα στάδια στο πλαίσιο των στοχαστικών ανελίξεων. Μία Μαρκοβιανή διαδικασία (Markov Process - MP), μπορεί να περιγραφεί ως ένα ζεύγος $\mathcal{M} = \langle \mathcal{S}, \mathcal{P} \rangle$, όπου $\mathcal{S}$ είναι ένα πεπερασμένο σύνολο καταστάσεων $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ και $\mathcal{P}$ είναι ο $|\mathcal{S}| \times |\mathcal{S}|$ πίνακας μετάβασης καταστάσεων, τέτοιος ώστε,

$$[\mathcal{P}]_{ij} = \mathbb{P}[S_{t+1} = s_j | S_t = s_i]$$

όπου $[\cdot]_{ij}$ συμβολίζει το στοιχείο της i-οστής γραμμής και j-οστής στήλης του πίνακα $\mathcal{P}$[4]. Στη συνέχεια θα χρησιμοποιήσουμε τον συμβολισμό $\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$.

Μία Μαρκοβιανή διαδικασία με επιβραβεύσεις (Markov Reward Process - MRP), είναι μια Μαρκοβιανή διαδικασία στην οποία όμως κάθε κατάσταση παράγει ένα σήμα επιβράβευσης R, το οποίο δειγματοληπτείται από κάποια άγνωστη κατανομή. Το MRP μπορεί να περιγραφεί ως μία τετράδα $\mathcal{M} = \langle \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, όπου $\mathcal{R} : \mathcal{S} \mapsto \mathbb{R}$ είναι η συνάρτηση επιβράβευσης έτσι ώστε $\mathcal{R}_s = \mathcal{R}(s) = \mathbb{E}[R_{t+1} | S_t = s]$[5], και $\gamma \in [0, 1]$ είναι ο παράγοντας υποβάθμισης της επιβράβευσης, αντίστοιχα με τον παράγοντα που χρησιμοποιήσαμε στα προβλήματα μίας κατάστασης. Αν G_t είναι η τυχαία μεταβλητή που περιγράφει τις αθροιστικές επιβραβεύσεις από τη χρονική στιγμή t και έπειτα, τότε

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

δηλαδή οι επιβραβεύσεις που πρόκειται να ληφθούν στο μέλλον υποβαθμίζονται. Συγκεκριμένα, αν κάποια κατάσταση s παράγει την επιβράβευση R, αλλά αναμένεται να μεταβούμε σε αυτή την κατάσταση μετά από k χρονικά βήματα, τότε η εκτιμώμενη επιβράβευση από αυτή

[4]Αν $\mathbf{1} = [1, 1, \ldots, 1]^T$ είναι το διάνυσμα μονάδων, είναι ενδιαφέρον να δούμε ότι $\mathcal{P}\mathbf{1} = \mathbf{1}$, ότι δηλαδή το διάνυσμα μονάδων είναι το πρώτο ιδιοδιάνυσμα του πίνακα μετάβασης καταστάσεων, το οποίο μάλιστα αντιστοιχεί στη μεγαλύτερη ιδιοτιμή. Εάν και στην παρούσα εργασία δεν θα ασχοληθούμε με το φασματικό περιεχόμενο του πίνακα κατάστασης καθώς και των πολλών διαφορετικών προσεγγίσεων που μπορεί να προκύψουν από την φασματική διάσπαση αυτού, παρόλα αυτά φαίνεται να υπάρχει μία πολύ ενδιαφέρουσα διασύνδεση με Λαπλασιανή θεωρία γραφημάτων [39]

[5]Σε αυτό το σημείο θα πρέπει να αναφέρουμε μία σύγχυση που ενδεχομένως να υπάρχει στους συμβολισμούς. Σε ένα μεγάλο μέρος της βιβλιογραφίας θεωρείται ότι η επιβράβευση από μία κατάσταση λαμβάνεται την αμέσως επόμενη χρονική στιγμή, δηλαδή πριν τη μετάβαση στη νέα κατάσταση. Αυτή τη λογική ακολουθούμε και παραπάνω καθώς χρησιμοποιείται και στο [56]. Παρ' όλα αυτά ένα σημαντικό μέρος θεωρεί ότι η επιβράβευση από μία κατάσταση λαμβάνεται την ίδια χρονική στιγμή, ότι δηλαδή $\mathcal{R}_s = \mathbb{E}[R_t | S_t = s]$.

την κατάσταση θα είναι $\gamma^k R$. Αν $\gamma = 0$, τότε δεν λαμβάνουμε υπόψιν επιβραβεύσεις από μελλοντικές καταστάσεις (αυτό συμβαίνει σε κάθε βήμα) και η εκτίμηση του G_t λέγεται *μυωπική*, ενώ αντίθετα αν $\gamma = 1$, τότε η εκτίμηση θα λέγεται *διορατική*. Η εξήγηση για τη χρήση του παράγοντα υποβάθμισης γ μπορεί να έχει πολλές πτυχές. Αρχικά είναι μαθηματικά βολικό, καθώς μπορούμε να έχουμε συγκλίνουσες ακολουθίες και κλειστή μαθηματική περιγραφή για τις εκτιμήσεις των αθροιστικών επιβραβεύσεων. Κατά δεύτερον μπορεί να αποφευχθούν κυκλικού τύπου λύσεις σε προβλήματα εύρεσης μονοπατιών, όπου σκοπός είναι η επίλυση του προβλήματος στον ελάχιστο δυνατό χρόνο. Κάτι τέτοιο προκύπτει και από βιολογικές παρατηρήσεις, καθώς οι βιολογικοί οργανισμοί προτιμούν μία άμεση επιβράβευση από μία καθυστερημένη επιβράβευση. Η συνάρτηση αξίας καταστάσεων $V : \mathcal{S} \mapsto \mathbb{R}$ μίας Μαρκοβιανής διαδικασίας με επιβράβευση, δίνει τη μακρέως χρόνου αξία των καταστάσεων. Με βάση τον ορισμό,

Ορισμός 5.3. *Η συνάρτηση αξίας καταστάσεων $V(s)$ ενός MRP είναι η εκτίμηση για τις αθροιστικές επιβραβεύσεις που θα ληφθούν, ξεκινώντας από την κατάσταση s, ως*

$$V(s) = \mathbb{E}[G_t|S_t = s] = \mathbb{E}[R_{t+1} + \gamma G_{t+1}|S_t = s] = \cdots = \mathcal{R}_s + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} V(s')$$

η οποία είναι η εξίσωση του Bellman. Για πεπερασμένο σύνολο $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$, μπορούμε να γράψουμε σε διανυσματική μορφή την παραπάνω σχέση, ώστε $\mathbf{v} = \mathcal{R} + \gamma \mathcal{P}\mathbf{v}$, όπου συμβολίζοντας με $\mathrm{v}(i) = V(s_i)$, και για απλούστευση $\mathcal{R}_i \equiv \mathcal{R}_{s_i}$, έχουμε

$$\begin{bmatrix} \mathrm{v}(1) \\ \vdots \\ \mathrm{v}(n) \end{bmatrix} = \begin{bmatrix} \mathcal{R}_1 \\ \vdots \\ \mathcal{R}_n \end{bmatrix} + \gamma \begin{bmatrix} \mathcal{P}_{11} & \cdots & \mathcal{P}_{1n} \\ \vdots & & \vdots \\ \mathcal{P}_{n1} & \cdots & \mathcal{P}_{nn} \end{bmatrix} \begin{bmatrix} \mathrm{v}(1) \\ \vdots \\ \mathrm{v}(n) \end{bmatrix}$$

δηλαδή η εξίσωση του Bellman είναι μία γραμμική εξίσωση η οποία μπορεί να λυθεί αναλυτικά, ως $\mathbf{v} = (I - \gamma\mathcal{P})^{-1}\mathcal{R}$, με κόστος $O(n^3)$ κάτι το οποίο δυσχεραίνει την λύση για MRPs με πολλές καταστάσεις, και συνήθως η λύση προκύπτει από επαναληπτικές διαδικασίες, όπως Δυναμικό Προγραμματισμό, μέθοδο Monte-Carlo, ή τον βασικό αλγόριθμο Temporal-Difference learning. Επεκτείνοντας τώρα τις ιδέες στις Μαρκοβιανές διαδικασίες λήψης αποφάσεων (MDPs), ένα MDP μπορεί να περιγραφεί ως μία επαύξηση των Μαρκοβιανών διαδικασιών με επιβράβευση εμπλουτισμένη με ένα πεπερασμένο σύνολο δράσεων $\mathcal{A}$. Συνεπώς μπορεί να περιγραφεί ως μία πεντάδα $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, όπου ο $\mathcal{P}$ ο νέος πίνακας μετάβασης καταστάσεων μέσω δράσεων, και θα συμβολίζουμε $\mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s'|S_t = s, A_t = a]$, ενώ $\mathcal{R}$ είναι νέα συνάρτηση για την εκτίμηση των επιβραβεύσεων από κάποια κατάσταση δεδομένης της δράσης, ώστε $\mathcal{R}^a_s = \mathbb{E}[R_{t+1}|S_t = s, A_t = a]$.

Ακολουθώντας με όμοιο τρόπο τις περιγραφές του κεφαλαίου 2, μπορούμε να ορίσουμε εκ νέου την πολιτική π.

Ορισμός 5.4. *Η πολιτική π λήψης αποφάσεων, είναι μία κατανομή στον χώρο δράσεων δεδομένων των καταστάσεων,*

$$\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$$

Συνολικά οι παραδοχές που γίνονται [56], είναι ότι (α) η πολιτική καθορίζει πλήρως την συμπεριφορά μίας μηχανής λήψης αποφάσεων, (β) ότι οι πολιτικές (όπως φαίνεται και από τον ορισμό) καθορίζονται μόνο από την τρέχουσα κατάσταση και όχι από την ιστορία, και (γ) ότι οι πολιτικές είναι στατικές, δηλαδή $A_t \sim \pi(\cdot|S_t), \forall t > 0$. Μπορούμε πλέον να δούμε ότι δεδομένου ενός MDP $\mathcal{M}_{MDP} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ και μίας πολιτικής π, η ακολουθία καταστάσεων που προκύπτει $S_1, S_2, \ldots$ είναι μία Μαρκοβιανή διαδικασία $\mathcal{M}_{MP} = \langle \mathcal{S}, \mathcal{P}^\pi \rangle$, ενώ η ακολουθία καταστάσεων και επιβραβεύσεων $S_1, R_2, S_2, R_3, \ldots$ είναι μία Μαρκοβιανή διαδικασία με επιβράβευση $\mathcal{M}_{MRP} = \langle \mathcal{S}, \mathcal{P}^\pi, \mathcal{R}^\pi, \gamma \rangle$, όπου συνηθίζεται να συμβολίζονται εκ νέου τα $\mathcal{P}^\pi_{ss'}$ και $\mathcal{R}^\pi_s$, ως

$$\mathcal{P}^\pi_{ss'} = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}^a_{ss'} \quad , \quad \mathcal{R}^\pi_s = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{R}^a_s \quad ,$$

Με βάση την παραπάνω περιγραφή, μπορούμε τώρα να γράψουμε τον ορισμό για τη συνάρτηση αξίας καταστάσεων σε ένα MDP.

Ορισμός 5.5. *Η συνάρτηση αξίας καταστάσεων $V_\pi(s)$ ενός MDP, είναι η εκτίμιση για τις αθροιστικές επιβραβεύσεις που θα ληφθούν, ξεκινώντας από την κατάσταση s και ακολουθώντας την πολιτική π*

$$V_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s]$$

ενώ επιπλέον ενδιαφέρον έχει η συνάρτηση αξίας δράσεων $Q_\pi(s, a)$ η οποία απαντάει στο ερώτημα του πόσο καλή είναι μία δράση βρισκόμενοι σε μία κατάσταση. Με βάση τον ορισμό,

Ορισμός 5.6. *Η συνάρτηση αξίας δράσεων $Q_\pi(s, a)$ ενός MDP, είναι η εκτίμηση για τις αθροιστικές επιβραβεύσεις που θα ληφθούν, ξεκινώντας από την κατάσταση s, εκτελώντας τη δράση a, και έπειτα ακολουθώντας την πολιτική π*

$$Q_\pi(s, a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a]$$

Από τους παραπάνω ορισμούς προκύπτουν οι *υπό εκτίμηση εξισώσεις του Bellman*. Για τον υπολογισμό της συνάρτησης αξίας καταστάσεων, βρισκόμενοι στην κατάσταση s, η πιθανότητα μετάβασης σε κάποια νέα κατάσταση ορίζεται από την πολιτική $\pi(a|s)$. Συνεπώς εάν γνωρίζουμε τις αξίες των δράσεων $Q_\pi(s,a)$ από την κατάσταση αυτή, τότε μπορούμε να εκτιμήσουμε την αξία της κατάστασης ως τον γραμμικό συνδυασμό των αξιών των δράσεων, με βάρη που αντιστοιχούν στην πιθανότητα να ληφθεί η κάθε δράση. Αντίστοιχα για τη συνάρτηση αξίας δράσεων, μπορούμε να πούμε ότι η αξία της δράσης a στην κατάσταση s θα είναι ίση με το άθροισμα της αναμενόμενης στιγμιαίας επιβράβευσης και της αξίας της κατάστασης προορισμού, υποβαθμισμένη με τον παράγοντα γ. Η κατάσταση προορισμού είναι αβέβαιη και καθορίζεται από τον πίνακα μετάβασης, οπότε και πάλι μπορεί να γραφεί σαν ένας κυρτός συνδυασμός χρησιμοποιώντας ως βάρη τις πιθανότητες μετάβασης. Αποτυπώνοντας όλα τα παραπάνω έχουμε τις υπο εκτίμηση εξισώσεις του Bellman,

$$V_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) Q_\pi(s,a) = \sum_{a\in\mathcal{A}} \pi(a|s) \left(\mathcal{R}_s^a + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}_{ss'}^a V_\pi(s') \right) \tag{5.1}$$

$$Q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}_{ss'}^a V_\pi(s') = \mathcal{R}_s^a + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}_{ss'}^a \sum_{a'\in\mathcal{A}} \pi(a'|s') Q_\pi(s',a') \tag{5.2}$$

όπου και πάλι η υπό εκτίμηση εξίσωση του Bellman για την συνάρτηση αξίας καταστάσεων μπορεί να γραφεί σε γραμμική μορφή $\mathbf{v}_\pi = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi \mathbf{v}_\pi$. Το ζητούμενο θα είναι να βρούμε τη βέλτιστη συνάρτηση αξίας καταστάσεων $V_*(s)$ και τη βέλτιστη συνάρτηση αξίας δράσεων $Q_*(s,a)$ για τις οποίες ισχύει

$$V_*(s) = \max_\pi V_\pi(s) \quad , \quad Q_*(s,a) = \max_\pi Q_\pi(s,a)$$

οι οποίες πρακτικά καθορίζουν τη βέλτιστη επίδοση που μπορεί κανείς να επιτύχει σε ένα MDP. Αντίστοιχα, όταν γνωρίζουμε τη βέλτιστη συνάρτηση αξίας (είτε καταστάσεων είτε δράσεων), τότε λέμε ότι το MDP είναι επιλυμένο. Η εξισώσεις (5.1, 5.2) γίνονται

$$V_*(s) = \max_a \left\{ \mathcal{R}_s^a + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}_{ss'}^a V_*(s') \right\} \quad , \quad Q_*(s,a) = \mathcal{R}_s^a + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}_{ss'}^a \max_{a'} Q_*(s',a')$$

οι οποίες είναι οι *εξισώσεις βέλτιστου του Bellman*[6]. Οι εξισώσεις αυτές είναι όπως φαίνεται

[6] Σαν συνέχεια του ερωτήματος που θέσαμε στο κεφάλαιο 1, οι βέλτιστες εξισώσεις του Bellman σε συνεχή χρόνο, δηλαδή για $t \to 0$, οδηγούν στην εξίσωση Hamilton-Jacobi-Bellman βελτίστου ελέγχου

μη γραμμικές, και γενικά δεν υπάρχει κλειστή λύση. Ωστόσο υπάρχουν αρκετοί επαναληπτικοί αλγόριθμοι που προσπαθούν να λύσουν το πρόβλημα της εύρεσης της βέλτιστης συνάρτησης, όπως Value Iteration, Policy Iteration, Q-learning, Sarsa.

5.2 Επίλυση MDPs με Δυναμικό Προγραμματισμό

Ως υπενθύμιση ο δυναμικός προγραμματισμός είναι στην ουσία του μία μέθοδος κατά την οποία ένα σύνθετο πρόβλημα μπορεί να χωριστεί σε υποπροβλήματα, τα υποπροβλήματα αυτά να λυθούν και η γενική λύση να προκύψει από τον συνδυασμό των λύσεων. Μπορεί να γίνει χρήση δυναμικού προγραμματισμού σε MDPs γνωστής δομής, είτε για πρόβλεψη, όπου η είσοδος στο πρόβλημα θα είναι το $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ και έξοδος η συνάρτηση V_π, είτε για έλεγχο, όπου η είσοδος θα είναι το $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ και έξοδος η βέλτιστη συνάρτηση V_* και η βέλτιστη πολιτική π_*.

Επαναλληπτική αξιολόγηση πολιτικής

Ας υποθέσουμε ότι θέλουμε να αξιολογήσουμε μία πολιτική λήψης αποφάσεων π. Μία αριθμητική μέθοδος είναι να υπολογίσουμε επαναληπτικά τις αξίες των καταστάσεων έτσι ώστε σε κάθε επανάληψη $k+1$, για κάθε κατάσταση $s \in \mathcal{S}$ να αναβαθμίσουμε την αξία της κατάστασης $V_{k+1}(s)$ με χρήση την αξίας $V_k(s')$ κάθε διάδοχης κατάστασης s'. Θεωρώντας ότι οι ανανεώσεις γίνονται με σύγχρονο τρόπο, τότε $\mathbf{v}^{k+1} = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi \mathbf{v}^k$. Αποδεικνύεται ότι οι τιμές της συνάρτησης αξίας για $k \to \infty$ θα συγκλίνουν στην V_π.

Προσέγγιση βέλτιστης πολιτικής

Το ζητούμενο εδώ είναι να βρούμε τη βέλτιστη πολιτική π_* (έστω προσεγγιστικά), δηλαδή να επιλύσουμε το MDP. Αρχικά χρειάζεται να μπορούμε σε ένα βήμα να βελτιώσουμε κάθε πολιτική π (διαδικασία policy improvement). Για να το επιτύχουμε αυτό, μπορούμε αρχικά να αξιολογήσουμε την πολιτική αυτή υπολογίζοντας τη συνάρτηση αξίας καταστάσεων $V_\pi(s) = \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \ldots | S_t = s]$ χρησιμοποιώντας τη μέθοδο επαναληπτικής αξιολόγησης. Έχοντας τη συνάρτηση $V_\pi(s)$, μπορούμε τώρα να ανανεώσουμε την πολιτική με greedy τρόπο, δηλαδή από κάθε κατάσταση s να επιλέγουμε ντετερμινιστικά τη δράση που οδηγεί στην κατάσταση με τη μεγαλύτερη αξία. Λαμβάνοντας υπόψιν τη στοχαστικότητα των μεταβάσεων, εάν $\pi'(s)$ περιγράφει τη νέα πολιτική λήψης απόφασης από την κατάσταση s, τότε

$$\pi'(s) = \operatorname*{argmax}_a \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} V_\pi(s')$$

Η επαναληπτική διαδικασία κατά την οποία γίνεται εκτίμηση της τρέχουσας πολιτικής V_π (π.χ. με επαναληπτική αξιολόγηση πολιτικής) και στο επόμενο βήμα γίνεται βελτίωση αυτής (π.χ.

με ντετερμινιστικό τρόπο policy improvement) ονομάζεται προσέγγιση βέλτιστης πολιτικής (αποδεικνύεται ότι η πολιτική συγκλίνει βέλτιστη βέλτιστη π_* [56]).

Προσέγγιση βέλτιστης αξίας

Το επόμενο ερώτημα αφορά στο κατά πόσο είναι αναγκαία η σύγκλιση στην V_π κατά την επαναληπτική αξιολόγηση πολιτικής, ώστε η προσέγγιση βέλτιστης πολιτικής να συγκλίνει πράγματι στην π_*. Θα μπορούσε κανείς να θέσει ένα όριο ϵ, έτσι ώστε να σταματάει την επαναληπτική αξιολόγηση πολιτικής στην επανάληψη k για την οποία $\|\mathbf{v}^{k+1} - \mathbf{v}^k\| \leq \epsilon$, ή απλά να θέσει ένα όριο για το k. Αποδεικνύεται ότι ακόμα και για $k = 1$, η πολιτική θα συγκλίνει στη βέλτιστη, και η συνάρτηση αξίας στη V_*. Μάλιστα μπορούμε να αποφύγουμε το βήμα βελτίωσης της πολιτικής, και απλά να αναβαθμίζουμε σε κάθε βήμα τη συνάρτηση αξίας καταστάσεων. Αποδεικνύεται ότι η επαναληπτική σχέση

$$V_{k+1}(s) = \max_{a \in \mathcal{A}} \left\{ \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a V_k(s') \right\} \Leftrightarrow \mathbf{v}^{k+1} = \max_{a \in \mathcal{A}} \left\{ R^a + \gamma \mathcal{P}^a \mathbf{v}^k \right\}$$

οδηγεί στη βέλτιστη συνάρτηση αξίας καταστάσεων. Η διαδικασία αυτή ονομάζεται προσέγγιση βέλτιστης αξίας (value iteration).

Να αναφέρουμε εδώ ότι οι ανανεώσεις στις παραπάνω επαναληπτικές μεθόδους γίνεται σύγχρονα, παρ' όλα αυτά υπάρχουν και ασύγχρονες μέθοδοι δυναμικού προγραμματισμού (όπως prioritized sweeping) οι οποίες μπορούν να μειώσουν κατά πολύ την υπολογιστική πολυπλοκότητα.

5.3 Πρόβλεψη και Έλεγχος σε MDPs Άγνωστης Δομής

Για τις μεθόδους δυναμικού προγραμματισμού που αναφέραμε, είναι απαραίτητη η γνώση της δομής (μεταβάσεις, επιβραβεύσεις, κλπ) του MDP. Στην περίπτωση που το MDP είναι άγνωστο, δύο είναι τα βασικά προβλήματα που μας απασχολούν. Αρχικά, η εκτίμηση της συνάρτησης αξίας καταστάσεων βάσει δεδομένης πολιτικής (πρόβλεψη), και δεύτερον η εύρεση της βέλτιστης συνάρτησης αξίας καταστάσεων ή/και της βέλτιστης πολιτικής (έλεγχος). Εάν και υπάρχει πληθώρα μεθόδων και προσεγγίσεων, εδώ θα αναφερθούμε περιληπτικά στις βασικότερες τις οποίες θα χρειαστούμε και στη συνέχεια, και οι οποίες αφορούν πρόβλεψη και έλεγχο χωρίς δημιουργία μοντέλου (model-free).

Μέθοδοι Monte Carlo και Temporal Differences

Αρχικά ως προς τη συνάρτηση αξίας καταστάσεων δεδομένης πολιτικής, μπορεί κανείς να εφαρμόσει εξαντλητικές μεθόδους Monte Carlo (MC). Επαναλαμβάνοντας ένα μεγάλο πλήθος αλληλεπιδράσεων πάνω στο άγνωστο MDP, εάν s είναι η τρέχουσα θέση, a η δράση που επιλέχθηκε και r η επιβράβευση που παρατηρήθηκε, τότε μπορεί κανείς από τις παρατηρήσεις

(s, a, r) να κάνει μία εκτίμηση για την αξία της κάθε κατάστασης $V(s)$ και την αξία της κάθε δράσης σε κάθε κατάσταση $Q(s, a)$ από τους αριθμητικούς μέσους των επιβραβεύσεων. Το βασικό μειονέκτημα εδώ, είναι ότι δεν εκμεταλλευόμαστε την γνώση παρά μόνο όταν τελειώσει ο προκαθορισμένος αριθμός επαναλήψεων, συνεπώς μία καλύτερη προσέγγιση είναι ο υπολογισμός του μέσου αθροιστικά ή ακόμα καλύτερα ο υπολογισμός του τρέχοντα μέσου. Θα θέλαμε λοιπόν να έχουμε έναν κανόνα ανανέωσης της μορφής $V(S_t) \leftarrow V(S_t) + \alpha(G_t - V(S_t))$, όπου όμως το G_t είναι άγνωστο καθώς εξαρτάται από τις μελλοντικές επιβραβεύσεις. Παρ' όλα αυτά μπορούμε να χρησιμοποιήσουμε τη λογική bootstraping έτσι ώστε να έχουμε μία καλή εκτίμηση με χρήση της υπάρχουσας γνώσης έτσι ώστε $G_t \approx R_{t+1} + \gamma V(S_{t+1})$. Αυτή η λογική οδηγεί στον βασικό αλγόριθμο TD(0). Εάν έχουμε την τετράδα παρατήρησης (s, a, r, s'), όπου s' η κατάσταση στην οποία βρεθήκαμε μετά τη δράση a από την κατάσταση s, και r η στιγμιαία επιβράβευση, τότε

$$V(s) \leftarrow V(s) + \alpha(r + \gamma V(s') - V(s))$$

όπου $r + \gamma V(s')$ είναι ο TD *στόχος*, ενώ $\delta = r + \gamma V(s') - V(s)$ είναι το TD *σφάλμα*. Τα πλεονεκτήματα του αλγορίθμου TD έναντι της μεθόδου MC είναι αρχικά ότι στον TD η μάθηση ξεκινάει από τα πρώτα βήματα, ενώ με MC οι ανανεώσεις γίνονται μετά το τέλος κάθε επεισοδίου[7]. Επίσης ο TD μπορεί να μάθει σε συνεχή περιβάλλοντα όπου δεν υπάρχει κατάσταση τερματισμού (non-terminating), ενώ ο MC λειτουργεί μόνο σε επεισοδιακά περιβάλλοντα (episodic - terminating). Εδώ δημιουργείται και ένα δίλημμα μεταξύ μεροληψίας και διασποράς (bias - variance). Για παράδειγμα, είναι προφανές (από τον ορισμό) ότι το G_t είναι μία αμερόληπτη εκτίμηση για το $V_\pi(S_t)$. Παρόλα αυτά η διασπορά του G_t είναι μεγάλη, καθώς οι εκβάσεις στην αλληλουχία των διαδοχικών καταστάσεων (και συνεπώς των επιβραβεύσεων) μπορεί να είναι τελείως διαφορετικές. Από την άλλη, ο TD στόχος $R_{t+1} + \gamma V(S_{t+1})$ δεν είναι αμερόληπτη εκτίμηση για το $V_\pi(S_t)$, ωστόσο η διασπορά είναι πολύ μικρότερη.

Ίχνη Δικαιοδοσίας - Eligibility Traces

Επεκτείνοντας τον TD(0), ένα εύλογο ερώτημα είναι γιατί να χρησιμοποιηθεί η τεχνική bootstrapping μετά το πρώτο βήμα και όχι μετά το δεύτερο, χρησιμοποιώντας ως στόχο το $G_t^{(2)} \approx R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2})$. Με το ίδιο σκεπτικό μπορεί να χρησιμοποιηθεί μετά το n-οστό βήμα υπολογίζοντας αντίστοιχα το $G_t^{(n)}$ και ανανεώνοντας την αξία της κάθε κατάστασης μετά από n-βήματα. Ακόμα καλύτερα, μπορούμε να χρησιμοποιήσουμε ως στόχο έναν γραμμικό συνδυασμό από τα διαφορετικά $G_t^{(n)}$, για $n = 1, 2, \ldots$, κάτι που τελικά οδηγεί στον αλγόριθμο TD(λ). Ο TD στόχος στον κανόνα ανανέωσης της συνάρτησης αξίας καταστάσεων, θα είναι

[7]Εδώ ως επεισόδιο αναφέρεται το σύνολο των εμπειριών μέχρι να οδηγηθούμε σε μία τελική κατάσταση τερματισμού.

$$G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{(n)}$$

με $\lambda \in [0,1]$. Παρ' όλα αυτά υπάρχει και εδώ το πρόβλημα της διορατικότητας, καθώς οι μελλοντικές επιβραβεύσεις είναι άγνωστες. Η λύση είναι η ιδέα της οπίσθιας πρόσβασης με χρήση *ιχνών δικαιοδοσίας* (eligibility traces), όπου σε κάθε βήμα t ανανεώνονται οι αξίες των καταστάσεων που βρίσκονται στο σύνολο της ιστορίας H_t. Για τον λόγο αυτό, σε κάθε κατάσταση s αντιστοιχεί ένας αριθμός $e(s)$, αρχικοποιημένος στο μηδέν, ο οποίος σε κάθε βήμα ανανεώνεται με βάση τον κανόνα $e(s) \leftarrow \gamma\lambda e(s) + \mathbb{I}\{S_t = s\}$. Έπειτα, υπολογίζοντας το TD σφάλμα δ, όλες οι καταστάσεις ανανεώνονται με χρήση του κανόνα $\mathbf{v} \leftarrow \mathbf{v} + \alpha\delta\mathbf{e}$, όπου $\mathbf{v}$ το διάνυσμα της συνάρτησης αξίας και $\boldsymbol{e}$ το διάνυσμα των ιχνών δικαιοδοσίας. Για $\lambda = 0$, είναι εμφανές ο αλγόριθμος γίνεται ο απλός TD(0).

Sarsa και Q-learning

Υποθέτουμε τώρα ότι έχουμε την παρατήρηση (s, a, r, s', a'), δηλαδή γνωρίζουμε ποια θα είναι η δράση a' από τη νέα κατάσταση s' (on-policy). Στον αλγόριθμο Sarsa, σε κάθε βήμα αναβαθμίζεται η συνάρτηση αξίας δράσεων ως για την κατάσταση s και τη δράση a ως

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma Q(s',a') - Q(s,a)$$

ο οποίος εγγυάται σύγκλιση στη βέλτιστη συνάρτηση Q_*, υπο την προϋπόθεση η πολιτική λήψης αποφάσεων να καταλλήγει ασυμπτωτικά σε greedy-ντετερμινιστική, και το βήμα μάθησης α να χρονομεταβλητό (ακολουθία Robbins-Monro) ώστε,

$$\sum_{t=1}^{\infty}\alpha_t = \infty \quad , \quad \sum_{t=1}^{\infty}\alpha_t^2 < \infty$$

Επιπλέον, ο αλγόριθμος Sarsa μπορεί να επεκταθεί στον Sarsa(λ) με τη χρήση ιχνών δικαιοδοσίας, παρόμοια με τον TD. Σε αυτόν τον αλγόριθμο υποθέσαμε ότι γνωρίζουμε ποια θα είναι η δράση a' από την κατάσταση s, και γι' αυτό λέμε ότι η αξιολόγηση και ο έλεγχος έγιναν *εντός-πολιτικής* (on-policy). Σε περίπτωση που δεν το γνωρίζουμε (π.χ η απόφαση δεν έχει ληφθεί ακόμα), ή στην περίπτωση που θέλουμε να αξιολογήσουμε κάποια άλλη πολιτική, τότε η μέθοδος αναβάθμισης και ελέγχου ονομάζεται *εκτός-πολιτικής* (off-policy). Ο πιο

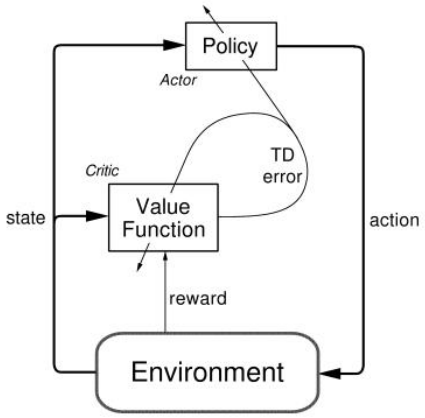
γνωστός αλγόριθμος εκτός-πολιτικής είναι ο αλγόριθμος Q-learning, κατά τον οποίο

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left(r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right)$$

και όπως φαίνεται ο στόχος $r + \gamma \max_{a'} Q(s',a')$ υπολογίζεται βάσει της greedy πολιτικής, ωστόσο οι δράσεις μπορούν να εκτελούνται βάσει της ϵ-greedy πολιτικής (π.χ. με κάποια από τις μεθόδους που αναφέρθηκαν στο κεφάλαιο 2 σε μία κατάσταση). Αποδεικνύεται [65] ότι ο αλγόριθμος Q-learning συγκλίνει στην βέλτιστη συνάρτηση αξίας δράσεων Q_*.

Μέθοδοι Δράστη - Κριτή

Όπως είναι πλέον σαφές, μπορεί να επιλέξει κανείς μεταξύ της εύρεσης της συνάρτησης αξίας καταστάσεων με δεδομένη κάποια πολιτική (π.χ. ϵ-greedy), είτε να εκτιμήσει την πολιτική μέσω της συνάρτησης αξίας δράσεων ανά κατάσταση. Θα έλεγε κανείς ότι και τα δύο αυτά είναι ισοδύναμα λόγω των εξισώσεων του Bellman. Όμως στις περιπτώσεις μάθησης χωρίς τη δημιουργία μοντέλου (τις οποίες μελετούμε), δεν γνωρίζουμε τις πιθανότητες μετάβασης. Με άλλα λόγια δεν γνωρίζουμε τη δομή του MDP, απλά εκτιμούμε τις συναρτήσεις αξίας (καταστάσεων-δράσεων) πάνω σε αυτό το άγνωστο MDP. Μία συνδυαστική μέθοδος είναι η μέθοδος δράστη κριτή, κατά την οποία εκτιμάται και η πολιτική αλλά και η συνάρτηση αξίας καταστάσεων. Ο δράστης (actor) χρησιμοποιεί την πολιτική για τη λήψη της απόφασης a από την κατάσταση s, ο κριτής (critic) λαμβάνει την επιβράβευση r και τη νέα κατάσταση s', υπολογίζει το TD σφάλμα δ και ανανεώνει τη συνάρτηση αξίας καταστάσεων, ενώ παράλληλα στέλνει το σφάλμα δ και στον δράστη για ανανέωση της πολιτικής (εικόνα 5.1).



Σχήμα 5.1: Διάγραμμα ενισχυτικής μάθησης με δράστη-κριτή (Πηγή [56]).

Κεφάλαιο 6

Προσαρμοστικός Αλγόριθμος Ενισχυτικής Μάθησης Παραμετροποιημένων Δράσεων με Δυναμική Εξερεύνηση ανά Κατάσταση

6.1 Εξειδικευμένη Εξερεύνηση ανά Κατάσταση

Όπως έχει γίνει ήδη αντιληπτό, η εξερεύνηση του χώρου κατάστασης και του χώρου δράσεων κατά τη διαδικασία μάθησης οφείλει να γίνει με προσεκτικά σχεδιασμένο τρόπο, έτσι ώστε να αποφευχθεί η σύγκλιση σε υποβέλτιστες πολιτικές λόγω κακών η μεροληπτικών αρχικοποιήσεων. Η στρατηγική σταδιακής μείωσης της εξερεύνησης (η οποία αναφέρθηκε στο κεφάλαιο 3 για τους χώρους μίας κατάστασης) μπορεί να επιφέρει πολύ κακές επιδόσεις στα δυναμικά περιβάλλοντα. Επιπρόσθετα, όταν η δυναμική των στατιστικών του περιβάλλοντος δεν περιορίζεται από αδρανειακούς κανόνες (και δεν είναι ανιχνεύσιμες συνολικά), θα πρέπει η μάθηση να συμπεριλαμβάνει την ικανότητα προσαρμοστικότητας. Στο [61] (το οποίο περιγράφεται επίσης στο κεφάλαιο 4), διερευνήσαμε ιδέες από το [15] σε χώρους μίας κατάστασης, ενώ στο [30] επεκτείναμε τις ιδέες αυτές για ανάπτυξη προσαρμοστικότητας σε παραμετροποιημένους χώρους δράσης και περισσότερες καταστάσεις, κάτω από το πλαίσιο της ενισχυτικής μάθησης.

Η βασική ιδέα που ακολουθήσαμε (και θα ακολουθηθεί και στη συνέχεια), είναι η δυναμική ρύθμιση της αβεβαιότητας των βέλτιστων παραμέτρων, ενώ ταυτόχρονα η εξερευνητική πολιτική καθορίζει ποια θα είναι η διακριτή δράση προς επιλογή. Αυτή η ιδέα μπορεί να περιγράφει και μία ιεραρχικού τύπου διαμέριση του πολύ υψηλών διαστάσεων συνεχή χώρου δράσεων, έτσι ώστε οι διαμερίσεις αυτές να περιγράφουν διακριτές δράσεις, μειώνοντας έτσι

την πολυπλοκότητα της αναζήτησης σε μικρότερες επιμέρους περιοχές[8]. Παρ΄ όλα αυτά, στο [29], επικεντρωθήκαμε στην αξιολόγηση του αλγορίθμου σε προβλήματα μίας κατάστασης, αν και ο φορμαλισμός ισχύει για πολλές καταστάσεις. Στη συνέχεια γενικεύουμε τον φορμαλισμό αυτό, όχι μόνο ενσωματώνοντας τη δυνατότητα για μοντελοποίηση με περισσότερες από μία παραμέτρους ανά δράση, αλλά χρησιμοποιώντας γραμμικούς χάρτες για την εύρεση της συνάρτησης αξίας καταστάσεων και δράσεων που περιγράφει το κάθε περιβάλλον. Επίσης, ο σκοπός μας είναι μια πιο μεθοδική προσέγγιση σε προβλήματα στα οποία κάποιες περιοχές του χώρου κατάστασης είναι στατικές ενώ άλλες πιο "ρευστές" (με μεγαλύτερη μεταβλητότητα). Σαν απλοϊκό παράδειγμα, ένας εκπαιδευτικός επιθυμεί να διδάξει τις μαθηματικές πράξεις της πρόσθεσης, αφαίρεσης, πολλαπλασιασμού και διαίρεσης σε μαθητές μικρής ηλικίας, έχοντας στο νου μία πληθώρα παραδειγμάτων και μεθόδων. Ο εκπαιδευτικός μπορεί να επιλέξει πόση ώρα θα αφιερώσει για τη διδασκαλία του κάθε τελεστή. Μπορεί επίσης να επιλέξει το πόσα παραδείγματα θα χρησιμοποιήσει, πόση εξάσκηση θα γίνει, ή ακόμα και πόσο περιγραφικός και επεξηγηματικός θα είναι. Ο στόχος θα είναι να ολοκληρώσει τη συνολική διδακτέα ύλη σε προκαθορισμένο χρονικό διάστημα, επιτυγχάνοντας την προσοχή των μαθητών του και κατανόηση όλων των μεθόδων. Ένας εκπαιδευτικός με πολλά χρόνια εμπειρίας, πιθανότατα έχει βελτιστοποιήσει και σταθεροποιήσει τις μεθόδους που χρησιμοποιεί για τη διδασκαλία της πρόσθεσης. Παρ΄ όλα αυτά ίσως είναι πιο αβέβαιος για τη διδασκαλία της διαίρεσης, συγκλίνοντας σε διαφορετική μέθοδο για κάθε τάξη σε κάθε σχολικό έτος, και χρησιμοποιώντας ως ανάδραση τα επίπεδα προσοχής και συμμετοχής των μαθητών. Επιστρέφοντας στη γενική ιδέα, η χρήση ενός και μόνο γενικού επίπεδου εξερεύνησης μπορεί να είχε ως αποτέλεσμα την άσκοπη αναζήτηση σε στατικά τμήματα του χώρου κατάστασης στα οποία η εξερεύνηση αυτή δεν ήταν απαραίτητη, αυξάνοντας τη μεταμέλεια των δράσεων.

Για την ανάπτυξη εξειδικευμένης εξερεύνησης ανά κατάσταση, αντικαθιστούμε τους καθολικούς μέσους βραχέως χρόνου $\bar{r}_t$ και μακρέως χρόνου $\bar{\bar{r}}_t$, με τη συνάρτηση αξίας κατάστασης $V(s)$ και $\bar{V}(s)$ αντίστοιχα, για κάθε κατάσταση ξεχωριστά. Για να γίνει κατανοητό, όταν η μηχανή μάθησης βρίσκεται σε κάποια κατάσταση s, εκτελεί κάποια δράση a με διάνυσμα παραμέτρων $\boldsymbol{\theta}$, παρατηρεί την επιβράβευση r και τη νέα κατάσταση s', ανανεώνουμε τη συνάρτηση αξίας $V(s)$ και τον τρέχων μέσο αυτής ως $\bar{V}(s) \leftarrow \bar{V}(s) + \alpha_V(V(s) - \bar{V}(s))$, ενώ χρησιμοποιούμε τη διαφορά τους $\bar{\delta}_V = V(s) - \bar{V}(s)$ για τη ρύθμιση του κλάσματος εξερεύνησης-αξιοποίησης για την κατάσταση s. Η δυναμική ρύθμιση των εξερευνητικών δράσεων μπορεί να δημιουργήσει προβλήματα αστάθειας, καθώς είναι πιθανή η ύπαρξη ενός θετικού κύκλου ανάδρασης. Μια μικρή εσφαλμένη αύξηση των εξερευνητικών δράσεων που μπορεί να προκύψει λόγω της στοχαστικότητας των επιβραβεύσεων και όχι λόγω αλλαγής του περιβάλλοντος, θα επιφέρει ακόμα χειρότερη μέση επίδοση βραχέως χρόνου. Όμως αυτό με τη σειρά του θα αυξήσει εκ νέου την εξερεύνηση και έτσι η επίδοση θα συνεχίσει να μειώνεται. Όπως περιγράφεται και κατά την αξιολόγηση του υβριδικού αλγορίθμου μετα-μάθησης για μη-στατικά περιβάλλοντα μίας κατάστασης στο κεφάλαιο κεφάλαιο 4 (αντίστοιχα στο [61]), περιπτώσεις

[8]Ενδεχομένως οι περιοχές αυτές μπορεί να έχουν και χαμηλότερη διάσταση, καθώς οι γενικευμένες δράσεις (π.χ δείχνω, πιάνω, περπατάω) θα χαρακτηρίζονται από μονοπάτια πάνω σε επιμέρους τοπολογικά πολύπτυχα (manifolds) του συνολικού χώρου δράσεων, με κάθε επιφύλαξη της δήλωσης αυτής.

στις οποίες μία υποβέλτιστη δράση γίνεται βέλτιστη, χωρίς ωστόσο να αλλάξουν (προς το χειρότερο) οι στατιστικές επιβράβευσης της βέλτιστης δράσης ως εκείνη τη στιγμή, μπορεί να μην ανιχνευθούν. Μία στρατηγική θα ήταν να υπάρχει ένα κάτω φράγμα στα επίπεδα εξερεύνησης, με κόστος να μην επιτευχθεί ποτέ επίδοση μηδενικής μεταμέλειας σε μεγάλες περιόδους στατικότητας του περιβάλλοντος [19].

6.2 Περιγραφή Προσαρμοστικού Αλγόριθμου

Θεωρούμε ένα διακριτό και πεπερασμένο σύνολο καταστάσεων $S = \{s_1, s_2, \ldots, s_k\}$, όπου η κάθε κατάσταση $s \in S$ μπορεί να αναπαρασταθεί από ένα m-διάστατο διάνυσμα κατάστασης $\boldsymbol{\phi}(s)$. Ο χώρος παραμετροποιημένων δράσεων μπορεί να περιγραφεί από ένα διακριτό και πεπερασμένο σύνολο $A_d = \{a_1, a_2, \ldots, a_n\}$, όπου κάθε $a \in A_d$ αντιστοιχεί σε μία διακριτή γενικευμένη δράση. Ακολουθώντας το πλαίσιο της ενισχυτικής μάθησης παραμετροποιημένων δράσεων [41], κάθε δράση μπορεί να περιγράφεται από ένα πλήθος m_a συνεχών παραμέτρων, ή απλά από ένα m_a-διάστατο διάνυσμα $\boldsymbol{\theta}^a \in \mathbb{R}^{m_a}$. Χρησιμοποιώντας τους παραπάνω συμβολισμούς, ο συνολικός χώρος δράσεων A μπορεί να γραφεί ως

$$A = \bigcup_{a \in A_d} \{(a, \boldsymbol{\theta}^a) | \boldsymbol{\theta}^a \in \mathbb{R}^{m_a}\}$$

Για την ανανέωση των αξιών των δράσεων θα χρησιμοποιήσουμε έναν απλό αλγόριθμο Q-learning, ο οποίος ωστόσο θα μπορούσε να αντικατασταθεί με οποιονδήποτε πιο σύνθετο (όπως $TD(\lambda)$ ή $Q(\lambda)$) ανάλογα με τις απαιτήσεις του προβλήματος.

Υποθέτουμε ότι η μηχανή μάθησης βρίσκεται στην κατάσταση s και επιλέγει μία δράση a με διάνυσμα παραμέτρων $\boldsymbol{\theta}^a$, ενώ το περιβάλλον επιστρέφει την επιβράβευση r και η νέα κατάσταση είναι s'. Σε αυτή τη φάση δεν εξηγούμε τον τρόπο που γίνεται η επιλογή της δράσης a και του διανύσματος παραμέτρων $\boldsymbol{\theta}^a$ (θα εξηγηθεί στη συνέχεια). Έχοντας λοιπόν διαθέσιμη την πεντάδα $(s, a, \boldsymbol{\theta}^a, r, s')$, μπορεί να υπολογιστεί το σφάλμα πρόβλεψης επιβράβευσης

$$\delta_Q = r + \gamma \max_{a_j} Q(s', a_j) - Q(s, a) \tag{6.1}$$

όπου γ είναι ο παράγοντας υποβάθμισης της επιβράβευσης. Το διάνυσμα παραμέτρων $\boldsymbol{\theta}^a$ δεν λαμβάνεται υπόψιν στον υπολογισμό του παραπάνω σφάλματος καθώς ως μέτρο εκτίμησης της αξίας των διακριτών δράσεων χρησιμοποιούνται μόνο οι τιμές Q. Παρ' όλα αυτά, μπορούμε να ακολουθήσουμε μία πιο γενική προσέγγιση κατά την οποία οι τιμές αξίες των δράσεων Q δεν είναι άμεσα προσπελάσιμες. Για παράδειγμα, αν $\boldsymbol{\phi}(s)$ αναπαριστά το m-διάστατο διάνυσμα χαρακτηριστικών της κατάστασης s, μπορούμε να χρησιμοποιήσουμε ένα νευρωνικό δίκτυο με

είσοδο το διάνυσμα αυτό και εξόδους που να εκτιμούν τη συνάρτηση αξίας δράσεων. Σε αυτή την εργασία θα προβούμε σε γραμμική προσέγγιση των $Q(s,a)$ από το $\boldsymbol{\phi}(s)$, και για τον λόγο αυτό θα χρησιμοποιήσουμε ένα νευρωνικό δίκτυο με m εισόδους και $n = |A_d|$ εξόδους (όπου $|\cdot|$ συμβολίζει την πληθικότητα) και με γραμμική συνάρτηση ενεργοποίησης $f(x) = x$ για κάθε νευρώνα. Παρά ταύτα οι ίδιες ιδέες μπορούν να επεκταθούν σε πιο βαθιές αρχιτεκτονικές ή άλλου είδους μη γραμμικές προσεγγίσεις.

Εάν $\boldsymbol{W}$ συμβολίζει τον $m \times n$ πίνακα βαρών του δικτύου, όπου w_{ij} είναι το βάρος που συνδέει την i-οστή συνιστώσα $\phi_i(s)$ με την j-οστή έξοδο, η οποία αντιστοιχεί στην αξία της δράσης a_j βρικόμενοι στην κατάσταση s, μπορούμε να γράψουμε ότι

$$\left[\boldsymbol{W}^T\boldsymbol{\phi}(s)\right]_j = Q(s,a_j) \tag{6.2}$$

όπου $[\cdot]_j$ συμβολίζει την j-οστή συνιστώσα ενός διανύσματος. Συμβολίζοντας με j τον δείκτη της δράσης που εκτελέστηκε, ώστε $a \equiv a_j$, το σφάλμα πρόβλεψης της επιβράβευσης δ_Q της εξίσωσης (6.1) μπορεί να γραφεί ως

$$\delta_Q = r + \gamma \max_i [\boldsymbol{W}^T\boldsymbol{\phi}(s')]_i - [\boldsymbol{W}^T\boldsymbol{\phi}(s)]_j \tag{6.3}$$

ενώ τα βάρη w_{ij} για κάθε $i \in \{1,2,\ldots,m\}$ ενημερώνονται με τον απλό κανόνα ανατροφοδότησης του σφάλματος

$$w_{ij} \leftarrow w_{ij} + \alpha_Q \delta_Q \phi_i(s) \tag{6.4}$$

όπου a_Q είναι ένα επιθυμητό βήμα μάθησης της επιλογής μας. Οι εκτιμήσεις των παραμέτρων $\boldsymbol{\theta}^a$ κάθε δράσης a μπορεί να περιγραφούν από έναν πίνακα $\hat{\boldsymbol{\Theta}}^a$ μεγέθους $m_a \times k$, με $k = |S|$, έτσι ώστε η j-οστή στήλη του $\hat{\boldsymbol{\Theta}}^a$ αντιστοιχεί στην εκτίμηση $\hat{\boldsymbol{\theta}}^a = \mathbb{E}[\boldsymbol{\theta}^a_\star | S = s_j]$, δηλαδή την εκτίμηση για το βέλτιστο διάνυσμα παραμέτρων $\boldsymbol{\theta}^a_\star$, της δράσης a στην κατάσταση s_j. Για την ενημέρωση των εκτιμήσεων αυτών θα χρησιμοποιήσουμε έναν αλγόριθμο δράστη-κριτή [59]. Εάν $\mathbf{v}$ είναι το διάνυσμα βαρών του κριτή, η τιμή της συνάρτησης αξίας καταστάσεων για την κατάσταση s μπορεί να γραφεί ως $V(s) = \mathbf{v}^T\boldsymbol{\phi}(s)$ και το σφάλμα πρόβλεψης της επιβράβευσης δ_V μπορεί να υπολογιστεί ως

$$\delta_V = r + \gamma \mathbf{v}^T\boldsymbol{\phi}(s') - \mathbf{v}^T\boldsymbol{\phi}(s) \tag{6.5}$$

με βάση το οποίο γίνεται η ανανέωση του διανύσματος βαρών $\mathbf{v}$ σε κάθε επανάληψη. Χρησιμοποιώντας βήμα μάθησης α_C θα έχουμε

$$\mathbf{v} \leftarrow \mathbf{v} + \alpha_C \delta_V \boldsymbol{\phi}(s) \tag{6.6}$$

Εάν με $\bar{\mathbf{v}}$ συμβολίσουμε τα βάρη για τη γραμμική προσέγγιση της τρέχουσας μέσης συνάρτησης $\bar{V}(s)$, τότε για την ανανέωσή τους θα πρέπει να χρησιμοποιηθεί η νέα εκτίμηση της αξίας της κατάστασης s ως στόχος. Χρησιμοποιώντας ένα βήμα μάθησης α_V της επιλογής μας, θα έχουμε ότι

$$\bar{\mathbf{v}} \leftarrow \bar{\mathbf{v}} + \alpha_V(\mathbf{v}^T \boldsymbol{\phi}(s) - \bar{\mathbf{v}}^T \boldsymbol{\phi}(s))\boldsymbol{\phi}(s) \tag{6.7}$$

ενώ το σφάλμα $\bar{\delta}_V = V(s) - \bar{V}(s)$ μπορεί να υπολογιστεί ως η διαφορά μεταξύ των εξόδων των δύο δικτύων

$$\bar{\delta}_V = \mathbf{v}^T \boldsymbol{\phi}(s) - \bar{\mathbf{v}}^T \boldsymbol{\phi}(s) \tag{6.8}$$

Για την εκτίμηση του $\hat{\boldsymbol{\Theta}}^a$ χρησιμοποιούμε ένα δίκτυο με πίνακα βαρών $\boldsymbol{G}^a$ μεγέθους $m \times m_a$, ξεχωριστό για κάθε δράση a. Επιπρόσθετα, οι τιμές των παραμέτρων μπορούν να περιοριστούν ώστε $\|\boldsymbol{\theta}^a\|_\infty \leq \theta_{\max}$. Η αναζήτηση των παραμέτρων μπορεί (και θα πρέπει), να γίνει σε έναν συμμετρικό χώρο με κατάλληλους αφινικούς μετασχηματισμούς γι΄ αυτόν το σκοπό. Εάν $\mathcal{F}_\theta(\cdot)$ είναι μία τμηματικά γραμμική συνάρτηση ενεργοποίησης, αποκόπτοντας τιμές μικρότερες από $-\theta_{\max}$ και μεγαλύτερες από $+\theta_{\max}$, η εκτίμηση για το διάνυσμα παραμέτρων $\hat{\boldsymbol{\theta}}^a$ θα είναι $\mathcal{F}_\theta((\boldsymbol{G}^a)^T \boldsymbol{\phi}(s))$. Έχοντας επιλέξει παραμέτρους $\boldsymbol{\theta}^a$, μπορούμε να υπολογίσουμε το διάνυσμα μετατόπισης $\boldsymbol{e}$ ως

$$\boldsymbol{e} = \boldsymbol{\theta}^a - \hat{\boldsymbol{\theta}}^a = \boldsymbol{\theta}^a - \mathcal{F}_\theta((\boldsymbol{G}^a)^T \boldsymbol{\phi}(s)) \tag{6.9}$$

και με βάση αυτό μπορούμε να ανανεώσουμε τον πίνακα $\boldsymbol{G}^a$. Εάν το σφάλμα πρόβλεψης της επιβράβευσης δ_V είναι θετικό (ακολουθώντας τη λογική του αλγορίθμου CACLA από [59]),

τότε για $i = \{1, \ldots, m\}$ και $j = \{1, \ldots, m_a\}$ ανανεώνουμε τα g_{ij}^a με

$$g_{ij}^a \leftarrow g_{ij}^a + \alpha_A \delta_V e_j \phi_i(s) \tag{6.10}$$

όπου α_A ένα νέο βήμα μάθησης. Επιστρέφουμε τώρα στην αρχή της περιγραφής της διαδικασίας για επεξήγηση της πολιτικής λήψης δράσεων. Βρισκόμενοι στην κατάσταση s η επιλογή του ζεύγους δράσης-παραμέτρων $(a, \boldsymbol{\theta}^a) \in A$ προκύπτει δειγματοληπτώντας την από κοινού συνάρτηση κατανομής πιθανότητας διακριτών δράσεων και παραμέτρων, όπου

$$p(a, \boldsymbol{\theta}^a|s) = p(a|s)p(\boldsymbol{\theta}^a|s, a)$$

Για την επιλογή της διακριτής δράσης θα χρησιμοποιήσουμε τη softmax Boltzmann όπως και στο [30], με τη βασική διαφοροποίηση ότι η παράμετρος αντίστροφης θερμοκρασίας θα είναι διαφορετική για κάθε κατάσταση s, ώστε

$$p(a|s) = \frac{\exp\{\beta(s)Q(s,a)\}}{\sum_{a \in A_d} \exp\{\beta(s)Q(s,a)\}} \tag{6.11}$$

όπου οι αξίες Q προκύπτουν από την αντίστοιχη γραμμική προσέγγιση με χρήση του πίνακα $\boldsymbol{W}$. Το διάνυσμα παραμέτρων της δράσης a που θα επιλεχθεί δειγματοληπτείται από μία πολυμεταβλητή Gaussian κατανομή κεντραρισμένη στην τρέχουσα εκτίμηση $\hat{\boldsymbol{\theta}}^a = \mathcal{F}_\theta((\boldsymbol{G}^a)^T \boldsymbol{\phi}(s))$ του βέλτιστου διανύσματος παραμέτρων $\boldsymbol{\theta}_\star^a$ και διαγώνιο πίνακα συνδιασποράς $\boldsymbol{\Sigma}_s^a$, ώστε

$$p(\hat{\boldsymbol{\theta}}^a|s, a) = \frac{1}{(2\pi)^{m_a/2}|\boldsymbol{\Sigma}_s^a|^{1/2}} \exp\{-\tfrac{1}{2}(\boldsymbol{\theta}^a - \hat{\boldsymbol{\theta}}^a)^T (\boldsymbol{\Sigma}_s^a)^{-1}(\boldsymbol{\theta}^a - \hat{\boldsymbol{\theta}}^a)\} \tag{6.12}$$

Το διάνυσμα βαρών $\boldsymbol{b}$ για τη γραμμική προσέγγιση του $\beta(s)$ αλλά και για τα διαγώνια στοιχεία του $\boldsymbol{\Sigma}_s^a$ ανανεώνονται αμέσως μετά τον υπολογισμό του $\bar{\delta}_V$. Επίσης χρησιμοποιούμε μία τμηματικά γραμμική συνάρτηση ενεργοποίησης $\mathcal{F}_\beta(\cdot)$ για το $\beta(s)$, αποκόπτοντας τιμές κάτω του μηδενός (αρνητικές τιμές δε θα είχαν νόημα), και μεγαλύτερες από κάποια $\beta_{\max}$, ώστε να αποφύγουμε την άσκοπη αύξηση σε τιμές μεγαλύτερες από αυτές όπου ούτως ή άλλως επιτυγχάνονται ικανοποιητικά επίπεδα αξιοποίησης των δράσεων. Έτσι θα έχουμε

$$\beta(s) = \mathcal{F}_\beta(\boldsymbol{b}^T \boldsymbol{\phi}(s)) \tag{6.13}$$

ενώ ο πίνακας συνδιασποράς $\boldsymbol{\Sigma}_s^a$ θα είναι ο διαγώνιος $\sigma^2(s,a)\boldsymbol{I}_{m_a}$, έτσι ώστε τα $\sigma^2(s,a)$ είναι η κοινή διασπορά που χαρακτηρίζει την αβεβαιότητα κατά την επιλογή των παραμέτρων της δράσης a. Με τη χρήση ενός επιπλέον διανύσματος βαρών $\boldsymbol{s}_a$ για κάθε δράση, οι τυπικές αποκλίσεις $\sigma(s,a)$ θα προκύπτουν από αντίστοιχο δίκτυο έτσι ώστε

$$\sigma(s,a) = \mathcal{F}_\sigma(\boldsymbol{s}_a^T \boldsymbol{\phi}(s)) \tag{6.14}$$

όπου $\mathcal{F}(\cdot)$ είναι μία σιγμοειδής συνάρτηση ενεργοποίησης μετατοπισμένη ώστε να έχει εύρος $(\sigma_{\min}, \sigma_{\max})$. Για την ενημέρωση των βαρών $\boldsymbol{b}$ και $\boldsymbol{s}_a$, ακολουθούμε διαφορετικό κανόνα ανάλογα με το πρόσημο του $\bar{\delta}_V$. Αν $\bar{\delta}_V \geq 0$, τότε

$$\boldsymbol{b} \leftarrow \boldsymbol{b} + \mu_\beta \bar{\delta}_V (\beta_{\max} - \mathcal{F}_\beta(\boldsymbol{b}^T \boldsymbol{\phi}(s))) \boldsymbol{\phi}(s) \tag{6.15}$$

$$\boldsymbol{s}_a \leftarrow \boldsymbol{s}_a - \mu_\sigma \bar{\delta}_V \mathcal{G}(\boldsymbol{s}_a^T \boldsymbol{\phi}(s)) \boldsymbol{\phi}(s) \tag{6.16}$$

όπου μ_β και μ_σ είναι παράμετροι προς επιλογή και $\mathcal{G}(\cdot)$ είναι η απλή σιγμοειδής συνάρτηση $\mathcal{G}(x) = 1/(1+e^{-x})$. Μία διαφοροποίηση είναι ότι εδώ χρησιμοποιούμε το $\bar{\delta}_V$ αντί του δ_V που χρησιμοποιούμε για την ανανέωση του $\boldsymbol{G}_a$. Όταν το δ_V είναι αρνητικό, χρησιμοποιούμε ένα κατώφλι $\bar{\delta}_{thr}$ και εξετάζουμε δύο υποπεριπτώσεις. Αν $\bar{\delta}_{thr} < \bar{\delta}_V < 0$ τότε τα βάρη παραμένουν ως έχουν, ενώ αν $\bar{\delta}_V \leq \bar{\delta}_{thr}$ τα βάρη ανανεώνονται με βάση τους παρακάτω κανόνες.

$$\boldsymbol{b} \leftarrow \boldsymbol{b} + \mu_\beta \bar{\delta}_V \mathcal{F}_\beta(\boldsymbol{b}^T \boldsymbol{\phi}(s)) \boldsymbol{\phi}(s) \tag{6.17}$$

$$\boldsymbol{s}_a \leftarrow \boldsymbol{s}_a - \mu_\sigma \bar{\delta}_V (1 - \mathcal{G}(\boldsymbol{s}_a^T \boldsymbol{\phi}(s))) \boldsymbol{\phi}(s) \tag{6.18}$$

Η χρήση του κατωφλίου $\bar{\delta}_{thr}$ είναι απαραίτητη, κυρίως στα στοχαστικά περιβάλλοντα στα οποία παρουσιάζεται αβεβαιότητα σε μεγάλο βαθμό, έτσι ώστε να αποφευχθεί η πυροδότηση του κύκλου θετικής ανάδρασης που αναφέραμε στην αρχή λόγω της μικρής ελάττωσης της επίδοσης. Οι μικρές πτώσεις που μπορεί να παρατηρούνται τοπικά στην επίδοση μπορεί να είναι

αποτέλεσμα της στοχαστικής φύσης του περιβάλλοντος και όχι λόγω κάποιας αλλαγής των στατιστικών του. Εάν κάτι τέτοιο δεν ληφθεί υπόψιν, τότε υπάρχουν συχνά φαινόμενα στα οποία η μηχανή μάθησης αυξάνει απότομα τις εξερευνητικές δράσεις, σαν ένα είδος επανεκκίνησης της πολιτικής με την οποία λαμβάνει τις αποφάσεις. Για αυτόν ακριβώς τον λόγο θα μπορούσαμε να συσχετίσουμε το κατώφλι αυτό με την ευαισθησία ενός ανιχνευτή διακοπτόμενων αλλαγών του περιβάλλοντος, παρόμοια με τον ανιχνευτή στατιστικών Page-Hinkley που αναφέραμε στο κεφάλαιο 3 και χρησιμοποιείται από τον αλγόριθμο Adapt-EvE. Η ρύθμιση του κατωφλίου $\bar{\delta}_{thr}$ μπορεί να γίνει με τρόπο ώστε οι αλλαγές να ανιχνεύονται μόνο όταν η σχετική πτώση της επίδοσης είναι σημαντική.

Μία άλλη προσέγγιση που έχει ενδιαφέρον, αφορά μία ειδική κατηγορία εφαρμογών κατά την οποία υπάρχει σήμα επιβράβευσης από κάθε κατάσταση s. Κάτι τέτοιο συνήθως αληθεύει σε σενάρια αλληλεπίδρασης ανθρώπου-ρομπότ όπου ναι μεν υπάρχει ένας γενικός στόχος (π.χ. επίλυση ενός προβλήματος), αλλά επιπρόσθετα υπάρχουν μικρότερες τοπικές επιβραβεύσεις που θα πρέπει να ληφθούν υπόψιν (π.χ μεγιστοποίηση της προσοχής). Σε αυτές τις περιπτώσεις το σήμα $\bar{\delta}_V$ που χρησιμοποιούμε για την ανανέωση της εξερευνητικής πολιτικής μπορεί να αντικατασταθεί από τις τρέχουσες μέσες επιβραβεύσεις βραχέως και μακρέως χρόνου που λαμβάνονται από την εκάστοτε κατάσταση, έτσι ώστε

$$\bar{\delta}_V = V_m(s) - \bar{V}_m(s)$$

όπου η τιμή της συνάρτησης αξίας καταστάσεων $V_m(s)$ στην κατάσταση s αντιστοιχεί στην τιμή της τρέχουσας μέσης επιβράβευσης από την εν λόγω κατάσταση. Στα περιβάλλοντα μίας κατάστασης και στους αλγορίθμους MLB και MLB-KF όπως περιγράφηκαν στο κεφάλαιο 4, μετά την επιστροφή επιβράβευσης r από το περιβάλλον, ο τρέχον μέσος των επιβραβεύσεων ανανεώνεται ως $\bar{r} \leftarrow \bar{r} + \alpha_C(r - \bar{r})$. Στο πλαίσιο της ενισχυτικής μάθησης, μετά την παρατήρηση της πεντάδας $(s, a, \boldsymbol{\theta}^a, r, s')$ θα έχουμε ανανέωση του τρέχοντα μέσου των επιβραβεύσεων από την κατάσταση s ως $\bar{r}(s) \leftarrow \bar{r}(s) + \alpha_C(r - \bar{r}(s))$. Αυτός ο κανόνας ανανέωσης είναι ισοδύναμος με την ανανέωση της συνάρτησης αξίας καταστάσεων $V(s)$

$$V(s) \leftarrow V(s) + \alpha_C(r + \gamma V(s') - V(s))$$

χρησιμοποιώντας $\gamma = 0$. Σε αυτές τις περιπτώσεις συμβολίζουμε τη συνάρτηση αξίας καταστάσεων ως $V_m(s)$, λόγω του μυωπικού τρόπου υπολογισμού. Αντίστοιχα ο υπολογισμός του τρέχοντα μέσου μακρέως χρόνου ή δεύτερης τάξης $\bar{\bar{r}}(s)$, όπου $\bar{\bar{r}}(s) \leftarrow \bar{\bar{r}}(s) + \alpha_V(\bar{r}(s) - \bar{\bar{r}}(s))$ είναι ισοδύναμος με τον τρέχοντα μέσο $\bar{V}_m(s)$. Αντίστοιχα, για τη ρύθμιση των αβεβαιοτήτων των παραμέτρων μέσω των $\sigma(s, a)$ μπορούμε να χρησιμοποιήσουμε τη διαφορά $\bar{\delta}_Q$ ως τη δια-

Algorithm 2 Active State-Specific Parameterized Exploration

Choose parameters $\alpha_{\{Q,C,V,A\}}, \mu_{\{\beta,\sigma\}}, \gamma, \beta_{\max}, \bar{\delta}_{thr}$
Initialize $\boldsymbol{G}^a, \boldsymbol{W}, \mathbf{v}, \bar{\mathbf{v}}, \boldsymbol{b}, \boldsymbol{s}_a$
Observe the initial state s
while true **do**
  Estimate $\beta(s), \sigma(s,a), Q(s,a)$ with Eq. 6.13, 6.2, 6.14
  Select an action tuple $(a, \boldsymbol{\theta}^a)$ with Eq. 6.11, 6.12
  Observe the new state s' and the reward r
  Compute the errors $\delta_Q, \delta_V, \boldsymbol{e}$ with Eq. 6.3, 6.5, 6.9
  Update the weights $\boldsymbol{W}, \mathbf{v}, \bar{\mathbf{v}}$ with Eq. 6.4, 6.6, 6.7
  Compute the error $\bar{\delta}_V$ with Eq. 6.8
  if $\delta_V > 0$ **then**
    Update the weights $\boldsymbol{G}^a$ with Eq. 6.10
  end if
  if $\bar{\delta}_V > 0$ **then**
    Update the weights $\boldsymbol{b}, \boldsymbol{s}_a$ with Eq. 6.15, 6.16
  else if $\bar{\delta}_V < \bar{\delta}_{thr}$ **then**
    Update the weights $\boldsymbol{b}, \boldsymbol{s}_a$ with Eq. 6.17, 6.18
  end if
  Set $s \leftarrow s'$
end while

φορά μεταξύ της μυωπικής αξίας $Q_m(s,a)$ της δράσης a στην κατάσταση s και του τρέχοντα μέσου αυτής $\bar{Q}_m(s,a)$, έτσι ώστε

$$\bar{\delta} = Q_m(s,a) - \bar{Q}_m(s,a)$$

Επιπρόσθετα, σε αυτές τις περιπτώσεις για την ενημέρωση των βαρών $\boldsymbol{b}$ στις εξισώσεις (6.15, 6.17) θα χρησιμοποιείται το $\bar{\delta}_V$, ενώ για την ενημέρωση των $\boldsymbol{s}_a$ στις εξισώσεις (6.16, 6.18) θα χρησιμοποιείται το $\bar{\delta}_Q$ αντί του $\bar{\delta}_V$. Η αναζήτηση των παραμέτρων μπορεί να γίνεται σε έναν κανονικοποιημένο χώρο και στη συνέχει να μεταφράζεται στις φυσικές τιμές (π.χ. στους επενεργητές του ρομπότ) με τους κατάλληλους αφινικούς μετασχηματισμούς $\boldsymbol{A\theta} + \boldsymbol{b}$, λαμβάνοντας υπόψην το εύρος κάθε παραμέτρου. Τα τελικά βήματα του αλγορίθμου μπορούν περιγράφονται στον Αλγόριθμο 2.

Στη συνέχεια παρουσιάζουμε ένα σύνολο από πειράματα τα οποία αναπτύχθηκαν για αξιολόγηση της επίδοσης του αλγορίθμου. Αρχικά γίνεται έλεγχος σε ένα περιβάλλον αναζήτησης μονοπατιού σε άγνωστο χάρτη και συνεχή χώρο κατάστασης, όπου επιδεικνύεται η προσαρμοστικότητα του αλγορίθμου σε μία σειρά από διακοπτόμενες αλλαγές του περι-

βάλλοντος, αλλάζοντας την τοπολογία του χάρτη και των διαθέσιμων μονοπατιών. Έπειτα γίνεται αξιολόγηση σε ένα περιβάλλον αριθμητικής προσομοίωσης αλληλεπίδρασης, όπου ο νέος αλγόριθμος ξεπερνά σε επίδοση την παλαιότερη έκδοση χωρίς εξειδικευμένη εξερεύνηση ανά κατάσταση κατά την ύπαρξη *καθολικών ή τοπικών*[9] αλλαγών του περιβάλλοντος. Έπειτα προσομοιώνουμε το πρόβλημα «ο πύργος του Ανόι», όπου το ρομπότ προσπαθεί να μάθει να επιλύει το πρόβλημα εκτελώντας ελάχιστο δυνατό αριθμό κινήσεων, ενώ παράλληλα ρυθμίζει κατάλληλα τις παραμέτρους που αφορούν την κίνηση, όπως την ταχύτητα των άκρων σε κάθε στάδιο για να μεγιστοποιήσει το ενδιαφέρον και την παρακολούθηση της επίλυσης από ένα παιδί.

6.3 Προσομοιώσεις σε Κλασσικά Προβλήματα Τεχνητής Νοημοσύνης

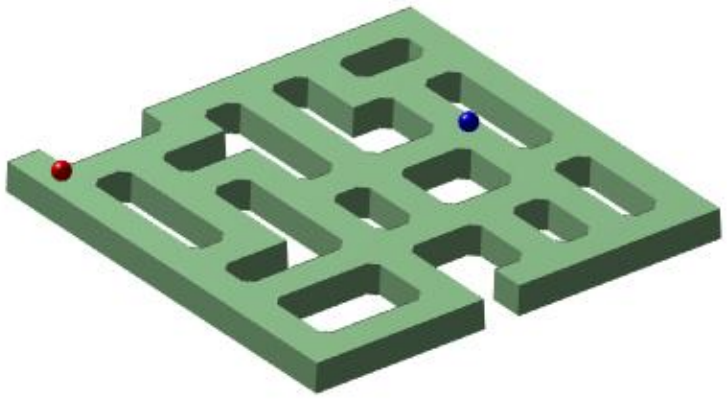
Παρόλο που οι βασικοί μας στόχοι είναι η επίδειξη των προσαρμοστικών δυνατοτήτων του αλγορίθμου σε σενάρια αλληλεπίδρασης ανθρώπου-ρομπότ, στη συνέχεια ελέγχουμε τις επιδόσεις του σε ένα πιο κλασσικό πρόβλημα τεχνητής νοημοσύνης, το οποίο αφορά την αναζήτηση μονοπατιού προς έναν επιθυμητό στόχο εντός λαβύρινθου άγνωστης τοπολογίας. Ένα τέτοιο πρόβλημα μπορεί να έχει αντιστοιχία σε εφαρμογές γενικότερου σκοπού (για παράδειγμα η αναζήτηση άγνωστου στόχου από το τελικό στοιχείο δράσης μίας ρομποτικής αλυσίδας με παράλληλη αποφυγή κινούμενων εμποδίων στο χώρο, αντιστοιχεί σε πρόβλημα αναζήτησης στον μη στάσιμο χάρτη του χώρου των αρθρώσεων).

6.3.1 Αναζήτηση Μονοπατιού σε Άγνωστο Χάρτη

Για τον έλεγχο του αλγορίθμου και της προσαρμοστικής του ικανότητας, θεωρούμε έναν υπερυψωμένο λαβύρινθο, χωρίς τοιχώματα, σε συνεχή χώρο 2 διαστάσεων όπως φαίνεται στο σχήμα 6.1. Πάνω στον λαβύρινθο βρίσκονται 2 μπάλες, μία κόκκινη και μία μπλε. Υποθέτουμε ότι ένα ρομπότ έχει τη δυνατότητα να λακτίσει την κόκκινη μπάλα στις 4 βασικές κατευθύνσεις (πάνω, κάτω, δεξιά, αριστερά). Ο στόχος είναι να κλωτσήσει την μπάλα τις λιγότερες δυνατές φορές ώστε αυτή να φτάσει και να χτυπήσει με τη σειρά της την μπλε μπάλα (στόχο). Η επιφάνεια του λαβύρινθου έχει επίσης κάποιον άγνωστο συντελεστή τριβής. Η κόκκινη μπάλα μπορεί μετά από μια προσπάθεια να πέσει από την υπερυψωμένη πλατφόρμα, και σε αυτή την περίπτωση επανατοποθετείται στην τελευταία θέση πριν την προσπάθεια αυτή. Επίσης η τοπολογία του λαβύρινθου (η οποία είναι άγνωστη στο ρομπότ), μπορεί να αλλάζει σε τυχαίες χρονικές στιγμές, είτε με την προσθήκη νέων μονοπατιών, την αφαίρεση κάποιου υπάρχοντος, αλλάζοντας την θέση της μπλε μπάλας-στόχου, ή αλλάζοντας τον συντελεστή τριβής της επιφάνειας. Ο χώρος διακριτών δράσεων θα είναι $A_d = \{up, down, left, right\}$ ενώ θα

[9] Σε αυτό το πλαίσιο οι καθολικές αλλαγές δεν αφορούν την αλλαγή των στατιστικών επιστροφής επιβράβευσης για κάθε δράση όπως στην περίπτωση των MAB, αλλά την αλλαγή της βέλτιστου ζεύγους δράσης-παραμέτρων για κάθε κατάσταση την ίδια χρονική στιγμή. Στις τοπικές αλλαγές υποθέτουμε ότι αυτό το ζεύγος μπορεί να αλλάζει μόνο σε κάποια κατάσταση κάθε χρονική στιγμή.

θεωρήσουμε μία παράμετρο ανά δράση $\theta^a \in (-3, 3)$ η οποία σχετίζεται άμεσα με τη δύναμη του λακτίσματος, και έμμεσα με την αρχική ταχύτητα v_0 με την οποία ξεκινάει την κίνηση η μπάλα, χωρίς να είναι γνωστές οι εξισώσεις κίνησης και η τριβή στο ρομπότ. Το ρομπότ μπορεί μόνο να εκτιμά την αρχική και τελική θέση της μπάλας σε σχέση με βάση κάποιο καθολικό σύστημα συντεταγμένων. Υποθέτουμε ότι υπάρχει κάποια μέγιστη ταχύτητα $v_{\max}$ που μπορεί να επιτευχθεί, η οποία αντιστοιχεί στη μέγιστη δύναμη του λακτίσματος που μπορεί να ασκήσει το ρομπότ. Στη συγκεκριμένη περίπτωση, η αναζήτηση των παραμέτρων θα γίνεται στο διάστημα $(-3, 3)$. Στη συνέχεια η τιμή αυτή θα καθορίζει τη δύναμη του λακτίσματος (εδώ δεν μας ενδιαφέρει η πραγματική εξάρτηση) ενώ μετά το λάκτισμα η μπάλα θα ξεκινάει την κίνηση με αρχική ταχύτητα $v_0 = (\theta + 3)v_{\max}$. Η παραπάνω προσέγγιση είναι βολική λόγω της Gaussian αναζήτησης που διεξάγουμε στον παραμετρικό χώρο, ενώ επίσης θα επιβάλουμε ένα άνω φράγμα $\sigma_{\max} = 3$, για την οποία η Gaussian συνάρτηση πυκνότητας πιθανότητας στο διάστημα $(-3, 3)$ είναι σχεδόν ομοιόμορφη. Με αυτόν τον τρόπο αφενός μεν δειγματοληπτούμε ομοιόμορφα όλον τον παραμετρικό χώρο όταν η αβεβαιότητα για την τιμή της παραμέτρου είναι υψηλή, αφετέρου περιορίζουμε την άσκοπη αύξηση της αβεβαιότητας πάνω από αυτό το κατώφλι ώστε να είναι εύκολη η επαναφορά της σε μικρότερες τιμές. Η ταχύτητα της κόκκινης μπάλας μειώνεται σε κάθε χρονικό βήμα με βάση τα χαρακτηριστικά του περιβάλλοντος. Για απλότητα κατά την προσομοίωση θα χρησιμοποιήσουμε τον κανόνα $v_{t+1} = (1 - f_c)v_t$, όπου $f_c = 0.05$ είναι μία παράμετρος συσχετιζόμενη με την τριβή του εδάφους, ενώ η μπάλα σταματάει αν $|v| < \epsilon$, όπου ϵ μία μικρή σταθερά (στις προσομοιώσεις χρησιμοποιήσαμε $\epsilon = 0.1$), και τότε το ρομπότ μπορεί να κάνει την επόμενη προσπάθεια.



Σχήμα 6.1: Διάταξη άγνωστου λαβύρινθου στο πρόβλημα αναζήτησης. Το ρομπότ μπορεί να λακτίσει την κόκκινη μπάλα στις 4 βασικές κατευθύνσεις, και λαμβάνει μία μοναδιαία επιβράβευση όταν την οδηγήσει στο να χτυπήσει την μπλε μπάλα. Ο λαβύριθνος είναι συνεχής και η κάθε μπάλα μπορεί να βρίσκεται οπουδήποτε στην επιφάνειά του.

Αναπαράσταση στο χώρο κατάστασης

Το ρομπότ δεν γνωρίζει την αρχική ταχύτητα της μπάλας ούτε τη δυναμική του περιβάλλοντος. Μπορεί μόνο να παρατηρεί την θέση $\boldsymbol{x}_b \in \mathbb{R}^2$ της μπάλας όταν αυτή είναι ακινητοποιημένη πάνω στην πλατφόρμα του λαβύρινθου με βάση ένα παγκόσμιο σύστημα αναφοράς, το οποίο χωρίς βλάβη της γενικότητας θα είναι τοποθετημένο στην πάνω αριστερή γωνία του λαβύρινθου. Η αναπαράσταση του διανύσματος κατάστασης $\boldsymbol{\phi}(s)$ μπορεί να γίνει με μία πληθώρα μεθόδων, όπως περιγράφεται στο [56]. Εδώ επιλέγουμε αρχικά τη χρήση πολλών ακτινικών συναρτήσεων βάσης (RBF) Gaussian τύπου, κάθε μία κεντραρισμένη σε σημείο $\boldsymbol{c}_i$ ενός $N \times M$ πλέγματος. Εάν $\boldsymbol{x}_b \in \mathbb{R}^2$ είναι η θέση της μπάλας στο 2Δ επίπεδο, τότε $\boldsymbol{d} \in \mathbb{R}^{N \times M}$ θα είναι ένα διάνυσμα ώστε

$$d_i = \exp\{-\|\boldsymbol{x}_b - \boldsymbol{c}_i\|^2 / 2\sigma^2\} \tag{6.19}$$

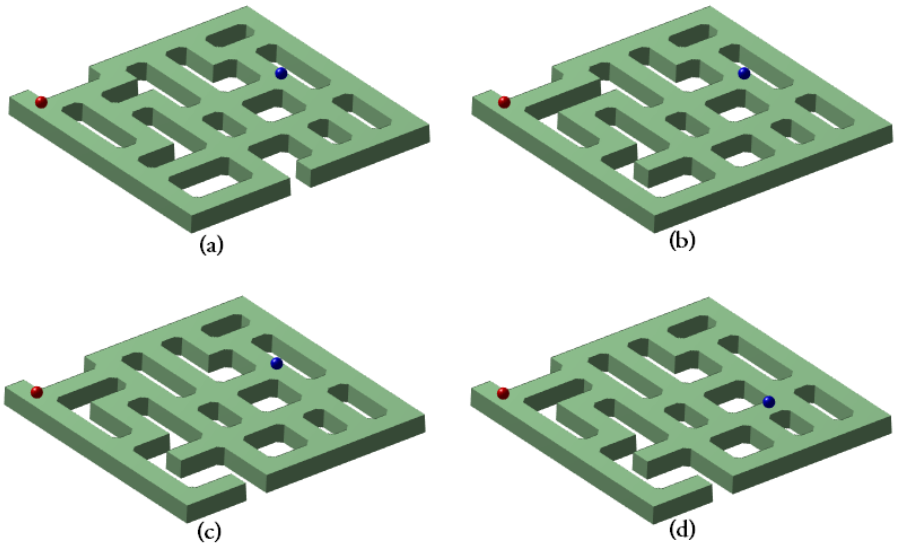
όπου σ είναι μία προκαθορισμένη παράμετρος με τιμή συγκρίσιμη με την απόσταση $\|\boldsymbol{c}_i - \boldsymbol{c}_j\|$ δύο γειτονικών σημείων $i \sim j$ στο πλέγμα. Με την κατάσταση s να περιγράφει τη θέση $\boldsymbol{x}_b$ της κόκκινης μπάλας πάνω στον λαβύρινθο, μία αρχική προσέγγιση για το διάνυσμα κατάστασης ήταν να δοκιμάσουμε $\boldsymbol{\phi}(s) = \boldsymbol{d}$, κάτι που παρ΄ όλα αυτά δεν είχε το ίδιο καλά τελικά αποτελέσματα όσο χρησιμοποιώντας διάνυσμα κατάστασης τύπου one-hot (ως one-hot εδώ εννοούμε την αναπαράσταση κατά την οποία μία και μόνο συνιστώσα του διανύσματος κατάστασης $\boldsymbol{\phi}(s)$ είναι μονάδα, ενώ όλες οι υπόλοιπες είναι μηδενικές), έτσι ώστε

$$\phi_i(s) = \mathbb{I}\{d_i = \|\boldsymbol{d}\|_\infty\}$$

Επίσης δοκιμάστηκε αναπαράσταση με τη μέθοδο coarse coding, κατά την οποία δημιουργούνται περισσότερα πλέγματα με διαγώνιες μετατοπίσεις των $\boldsymbol{c}_i$ και επαυξάνοντας το διάνυσμα κατάστασης ως την συνένωση των επιμέρους one-hot αναπαραστάσεων που προκύπτουν από το κάθε πλέγμα. Παρ΄ όλα αυτά, ενώ αυξήθηκε κατά πολύ η πολυπλοκότητα δεν υπήρχε κάποια αισθητή βελτίωση της επίδοσης που θα παρουσιαστεί στη συνέχεια.

Το πρόβλημα ξεκινάει όπως φαίνεται στο σχήμα 6.2a. Μετά από κάθε προσπάθεια, η μηχανή μάθησης αμείβεται με επιβράβευση $r = 1$ αν η κόκκινη μπάλα χτυπήσει την μπλε μπάλα, ή με μηδέν σε κάθε άλλη περίπτωση (ακόμα και αν πέσει από τον λαβύρινθο). Όταν το παιχνίδι τελειώσει επιτυχώς, οι δύο μπάλες τοποθετούνται στις αρχικές τους θέσεις, αποθηκεύεται ο αριθμός των προσπαθειών που χρειάστηκαν ώστε να γίνει αξιολόγηση της διαδικασίας[10] και το παιχνίδι ξανά αρχίζει. Σε περίπτωση που το παιχνίδι δεν έχει ολοκληρωθεί μετά από ένα άνω

[10] Σε όλες τις πιθανές διατάξεις που θα χρησιμοποιηθούν για τον λαβύρινθο, οι ελάχιστες δυνατές προσπάθειες που χρειάζονται για να ολοκληρωθεί επιτυχώς το παιχνίδι είναι τρεις.

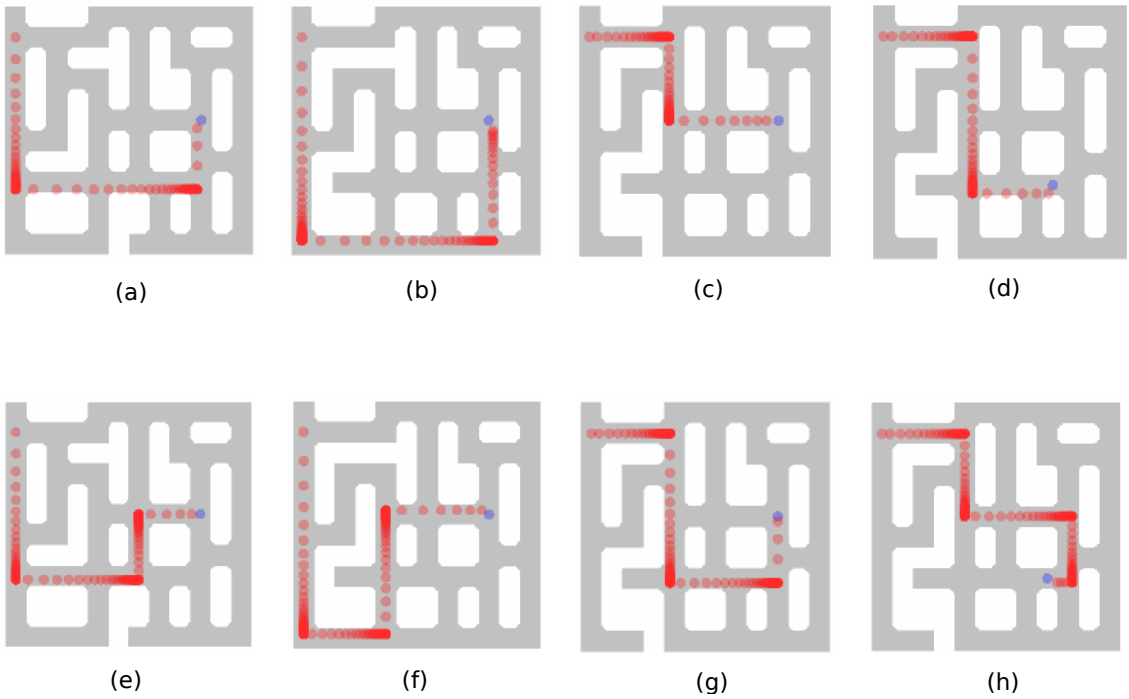


Σχήμα 6.2: Διαφορετικές διατάξεις του λαβύρινθου για το πρόβλημα αναζήτησης. (a): Αρχική διάταξη για τα παιχνίδια 1-5000. (b): Διάταξη για τα παιχνίδια 5001-10000. Το βέλτιστο μοναπάτι έχει αφαιρεθεί, ενώ έχει προστεθεί μία νέα διαδρομή έτσι ώστε ο βέλτιστος αριθμός προσπαθειών να παραμένει ο ίδιος. (c): Διάταξη για τα παιχνίδια 10001-15000. Το βέλτιστο μονοπάτι αποκόπτεται και δημιουργείται καινούργιο. (d): Διάταξη μετά το τρίτο σημείο αλλαγής, για τα παιχνίδια 15001-20000. Η θέση της μπλε μπάλας αλλάζει.

όριο προσπαθειών (χρησιμοποιήθηκε ως άνω όριο το 1000) το παιχνίδι τελειώνει ανεπιτυχώς και αρχίζει από την αρχή. Μετά από 5000 παιχνίδια, η τοπολογία του λαβύρινθου αλλάζει, αφαιρώντας κάποιο τμήμα που ήταν απαραίτητο για την ολοκλήρωση της διαδικασίας σε 3 προσπάθειες, και προσθέτοντας κάποιο άλλο τμήμα έτσι ώστε ο ελάχιστος δυνατός αριθμός προσπαθειών να παραμένει 3, όπως φαίνεται στο σχήμα 6.2b. Το ρομπότ δεν γνωρίζει τίποτα για την αλλαγή και συνεχίζει να έχει πρόσβαση μόνο στη θέση της κόκκινης μπάλας και να λαμβάνει επιβράβευση $r = 1$ όταν επιτύχει τον στόχο. Στο παιχνίδι 10000 η τοπολογία του λαβύρινθου αλλάζει και πάλι εμποδίζοντας εκ νέου το βέλτιστο μονοπάτι και δημιουργώντας καινούργιο όπως φαίνεται στο σχήμα 6.2c, ενώ στο παιχνίδι 15000 η τοποθεσία της μπλε μπάλας αλλάζει όπως φαίνεται στο σχήμα 6.2d. Στο παιχνίδι 20000 η προσομοίωση τελειώνει και ονομάζουμε όλη αυτή την εμπειρία ως υπερ-συνεδρία.

Ρύθμιση Υπερπαραμέτρων

Ένα από τα μειονεκτήματα του αλγορίθμου είναι ο μεγάλος αριθμός υπερπαραμέτρων που χρειάζεται να ρυθμιστούν. Χρησιμοποιώντας επιβραβεύσεις $r \in [0, 1]$, μπορούμε να θέσουμε τη μέγιστη τιμή $\beta_{\max}$ της αντίστροφης θερμοκρασίας αλλά και το κατώφλι ευαισθησίας $\bar{\delta}_{thr}$ χειροκίνητα. Αν και ομολογουμένως θα ήταν χρήσιμη μία αναλυτική προσέγγιση, κάτι τέτοιο είναι εκτός του φάσματος της παρούσας εργασίας καθώς επίσης δεν μπορούμε να εγγυηθούμε ότι μία τέτοια αναλυτική προσέγγιση είναι εφικτή. Στη συγκεκριμένη περίπτωση τα



Σχήμα 6.3: Εύρεση μονοπατιών σε μεταβλητό λαβύρινθο. (a): Ένα στιγμιότυπο παιχνιδιού στον 1ο λαβύρινθο, όπου το ρομπότ έχει μάθει τα βέλτιστα ζεύγη δράσης-παραμέτρων. Σε αυτό το στιγμιότυπο η κόκκινη μπάλα βρέθηκε πολύ κοντά το να πέσει από την πλατφόρμα. (b): Στιγμιότυπο παιχνιδιού για τον 2ο λαβύρινθο όπου το βέλτιστο μονοπάτι έχει βρεθεί. (c): Στιγμιότυπο παιχνιδιού για τον 3ο λαβίρυνθο όπου το βέλτιστο μονοπάτι έχει βρεθεί. (d): Στιγμιότυπο παιχνιδιού για τον 4ο λαβύρινθο, όπου το νβέλτιστο μονοπάτι έχει βρεθεί. (e): Στιγμιότυπο παιχνιδιού για τον 1ο λαβύρινθο, όπου το ρομπότ ακολούθησε ένα υποβέλτιστο μονοπάτι. (f): Στιγμιότυπο παιχνιδιού για τον 2ο λαβύρινθο, όπου το ρομπότ ακολούθησε ένα υποβέλτιστο μονοπάτι. (g): Στιγμιότυπο παιχνιδιού για τον 3ο λαβύρινθο, όπου το ρομπότ ακολούθησε ένα υποβέλτιστο μονοπάτι. (h): Στιγμιότυπο παιχνιδιού για τον 4ο λαβύρινθο, όπου το ρομπότ ακολούθησε ένα υποβέλτιστο μονοπάτι.

$\beta_{\max}, \bar{\delta}_V$ συμπεριλήφθηκαν στην αναζήτηση υπερπαραμέτρων, ενώ χρησιμοποιήσαμε σταθερό παράγοντα υποβάθμισης των επιβραβεύσεων $\gamma = 0.9$. Εφόσον η αναπαράσταση για τον χώρο κατάστασης ήταν η one-hot, η παράμετρος σ της Gaussian RBF για τον υπολογισμό των d_i στην εξίσωση 6.19 δεν παίζει κανένα ρόλο για τη δημιουργία του διανύσματος κατάστασης $\boldsymbol{\phi}(s)$. Πρακτικά οποιαδήποτε τιμή $\sigma \neq 0$ και να επιλέξουμε, το διάνυσμα $\boldsymbol{\phi}(s)$ θα έχει μονάδα στη συνιστώσα i που αντιστοιχεί στο σημείο του πλέγματος $\boldsymbol{c}_i$ που βρίσκεται κοντύτερα στην κόκκινη μπάλα, και μηδέν σε όλες τις άλλες συνιστώσες. Εδώ χρησιμοποιήσαμε $\sigma = 0.7$ και επιλέξαμε πλέγμα με $N = M = 14$, συνεπώς το διάνυσμα κατάστασης ήταν 196 διαστάσεων.

Αρχικά επιλέξαμε χειροκίνητα ένα σύνολο παραμέτρων μετά από απλή παρατήρηση της επίδοσης για διάφορες επιλογές, και το θεωρήσαμε ως το σύνολο αρχικοποίησης. Έπειτα αξιολογήσαμε την επίδοση της μηχανής μάθησης εκτιμώντας το *ποσοστό βελτιστότητας*, το

οποίο υπολογίστηκε ως το πηλίκο του αριθμού τον παιχνιδιών τα οποία ολοκληρώθηκαν στον ελάχιστο αριθμό προσπαθειών, προς το σύνολο των παιχνιδιών στην υπερ-συνεδρία (δηλαδή 25000). Στο σχήμα 6.3 φαίνονται τα βέλτιστα μονοπάτια (a, b, c, d) αλλά και κάποια υποβέλτιστα μονοπάτια (e, f, g, h) που ακολούθησε το ρομπότ κατά τη διαδικασία της μάθησης για τα αντίστοιχα 4 διαφορετικά περιβάλλοντα του σχήματος 6.2. Λόγω της αραιής αναπαράστασης του χώρου κατάστασης και της στοχαστικότητας των επιλογών (χρησιμοποιήσαμε $\sigma_{\min} = 0.05$), κάθε παιχνίδι είναι διαφορετικό και μοναδικό, ακόμα και όταν η πολιτική με την οποία η μηχανή μάθησης λαμβάνει αποφάσεις έχει σταθεροποιηθεί. Για παράδειγμα, στο σχήμα 6.3a, το βέλτιστο ζεύγος διακριτής δράσης και παραμέτρου $(a^\star, \theta^a_\star)$ έχει βρεθεί, παρά ταύτα σε κάποιες προσπάθειες η κόκκινη μπάλα μπορεί να πέσει εκτός της πλατφόρμας του λαβύρινθου, καθώς η ύπαρξη του $\sigma_{\min}$ στην επιλογή της παραμέτρου εισάγει στοχαστικότητα η οποία είναι ισοδύναμη με την ύπαρξη ενός παρατηρήσιμου Gaussian θορύβου σε μία ντετερμινιστικού τύπου πολιτική επιλογής παραμέτρου. Παρατηρήθηκαν επίσης περιπτώσεις κατά τις οποίες πραγματοποιήθηκαν δύο ή περισσότερα διαδοχικά λακτίσματα προς την ίδια κατεύθυνση, αντί ενός με καταλληλότερη επιλογή παραμέτρου (πιο δυνατό), όπως επίσης και περιπτώσεις όπου το χτύπημα ήταν πολύ δυνατό οδηγώντας την μπάλα σε σημείο πιο μακρυά από το βέλτιστο, ακολουθούμενο από χτύπημα προς την αντίθετη κατεύθυνση για την τοποθέτηση της μπάλας στην αρχή του κατάλληλου διαδρόμου.

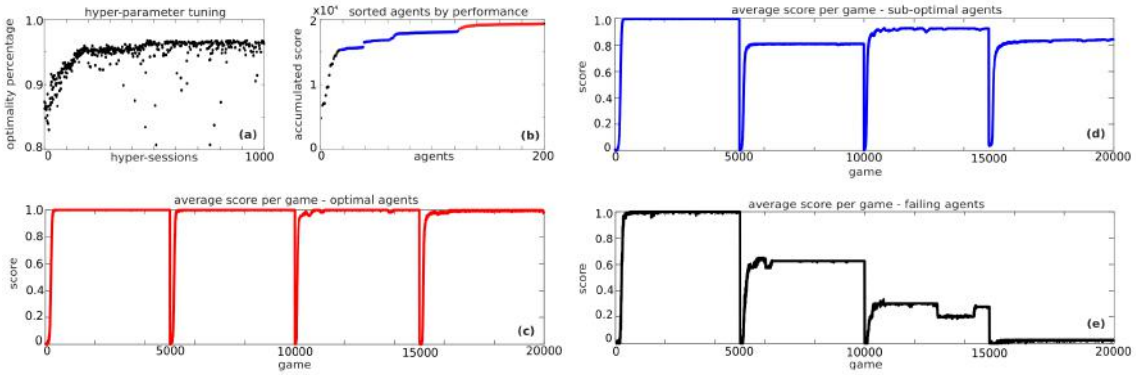
Για τη ρύθμιση των παραμέτρων πραγματοποιήθηκε μία τοπικά τυχαία ομοιόμορφη αναζήτηση εντός υπερκύβικής περιοχής κεντραρισμένης ανά το πέρας χρονικών διαστημάτων στο εμπειρικά τρέχων βέλτιστο σημείο του παραμετρικού χώρου, χρησιμοποιώντας ως μετρική επίδοσης το ποσοστό βελτιστότητας. Η αναζήτηση στον χώρο παραμέτρων πραγματοποιήθηκε για 1000 υπερ-συνεδρίες (συνολικά 2×10^7 παιχνίδια) και τα αποτελέσματα για κάθε συνεδρία φαίνονται στο σχήμα 6.4a (σε αυτές τις γραφικές δεν φαίνεται το ποσοστό βελτιστότητας για περιπτώσεις στις οποίες το ρομπότ κατέληξε στο να επιλέγει ένα υποβέλτιστο μονοπάτι). Οι παράμετροι που τελικώς επιλέχθηκαν και αντιστοιχούν στην υπερ-συνεδρία για την οποία επιτεύχθηκε η καλύτερη επίδοση φαίνονται στον πίνακα 6.1. Οι τιμές αυτές χρησιμοποιήθηκαν για την αξιολόγηση του αλγορίθμου όπως περιγράφεται στην επόμενη παράγραφο.

α_Q	α_A	α_C	α_V	μ_β	μ_σ	$\beta_{\max}$	$\bar{\delta}_{thr}$
0.142	0.433	0.196	0.005	0.205	1.387	16.2	-0.311

Πίνακας 6.1: Βέλτιστες τιμές παραμέτρων για το πρόβλημα αναζήτησης σε χάρτη, οι οποίες επιλέχθηκαν μετά από προσαρμοστική τυχαία αναζήτηση.

Αποτελέσματα

Χρησιμοποιώντας τις τιμές των παραμέτρων του πίνακα 6.1, επαναλάβαμε 200 υπερ-συνεδρίες αποθηκεύοντας τον αριθμό των προσπαθειών που χρειάστηκαν για να ολοκληρωθεί κάθε παιχνίδι. Μετά το τέλος κάθε υπερ-συνεδρίας η μνήμη του ρομπότ διαγράφεται και αρχικοποιείται



Σχήμα 6.4: Αποτελέσματα για το πρόβλημα αναζήτησης σε λαβύρινθο. (a): Προσαρμοστική ρύθμιση υπερ-παραμέτρων. Κάθε σημείο αντιστοιχεί στην επίδοση του αλγορίθμου για το επιλεγμένο σύνολο. (b): Η ταξινομημένες επιδόσεις των 200 εικονικών ρομπότ χρησιμοποιώντας το βέλτιστο εμπειρικό σύνολο υπερ-παραμέτρων. Το κόκκινο χρώμα αντιστοιχεί στα εικονικά ρομπότ με βέλτιστο τύπο προσαρμοστικότητας, το μπλε χρώμα στα ρομπότ υπο-βέλτιστου τύπου προσαρμοστικότητας, και το μαύρο χρώμα στα ρομπότ που απέτυχαν να προσαρμοστούν στις αλλαγές. (c): Μέσο σκορ ανά παιχνίδι, των ρομπότ βέλτιστου τύπου. (d): Μέσο σκορ ανά παιχνίδι, των ρομπότ υπο-βέλτιστου τύπου. (e): Μέσο σκορ ανά παιχνίδι, των ρομπότ που αποτυγχάνουν.

εκ νέου. Στη συνέχεια ταξινομήσαμε τα εικονικά ρομπότ με βάση την ακόλουθη λογική. Εάν ένα ρομπότ κατάφερε να ολοκληρώσει κάποιο παιχνίδι στον ελάχιστο αριθμό προσπαθειών, τότε προσθέταμε μία μονάδα στο συνολικό του σκορ. Για κάθε παραπάνω βήμα που εκτελούσε μειώναμε την τιμή αυτή κατά 0.25, συνεπώς το σκορ ήταν μηδενικό αν ολοκλήρωνε το παιχνίδι σε 4 ή και περισσότερα βήματα από τα ελάχιστα δυνατά. Τα αθροιστικά σκορ όλων των 200 εικονικών ρομπότ φαίνονται στο σχήμα 6.4b. Περίπου το 39% των ρομπότ (agents 128-200) όχι μόνο προσαρμόστηκαν σε κάθε αλλαγή του λαβύρινθου, αλλά βρήκαν το νέο βέλτιστο μονοπάτι σε κάθε περίπτωση ολοκληρώνοντας το παιχνίδι στον ελάχιστο αριθμό βημάτων. Το 56% (agents 11-127) προσαρμόστηκαν σε κάθε αλλαγή και ολοκλήρωσαν τα νέα παιχνίδια, αλλά κατά μέσο όρο με μια περισσότερη προσπάθεια για κάποια από τα περιβάλλοντα (το 2ο και το 4ο). Το υπόλοιπο 5% (agents 1-10) δεν προσαρμόστηκαν και μάλιστα τα επίπεδα προσαρμογής τους έπεφταν από αλλαγή σε αλλαγή. Η εικόνα 6.4(c,d,e) απεικονίζει το μέσο αθροιστικό σκορ για κάθε μια κατηγορία από αυτά τα ρομπότ, με κόκκινο μπλε και μαύρο χρώμα αντίστοιχα. Συνολικά τα αποτελέσματα αυτά είναι ενθαρρυντικά και επιδεικνύουν την προσαρμοστική ικανότητα του αλγορίθμου σε ένα αρκετά σύνθετο και πρωτότυπο πρόβλημα.

6.4 Προσομοιώσεις σε Περιβάλλοντα Αλληλεπίδρασης Ανθρώπου-Ρομπότ

Στις εφαρμογές αλληλεπίδρασης ανθρώπου-ρομπότ που μας ενδιαφέρουν, ο στόχος είναι η μεγιστοποίηση του ενδιαφέροντος και της προσοχής του ανθρώπου κατά τη διάρκεια μίας εργασίας που εκτελείται από το ρομπότ. Για όλα τα επόμενα προβλήματα, θα χρησιμοποιήσουμε το πλαίσιο που έχει αναπτυχθεί στο [30]. Αρκετές παλαιότερες έρευνες στην περιοχή αλληλεπίδρασης ανθρώπου-ρομπότ, έχουν δείξει ότι η ποσοτικοποίηση της ανθρώπινης προσοχής είναι χρήζουσας σημασίας για τη μέτρηση και την βελτιστοποίηση της ποιότητας της αλληλεπίδρασης [1]. Παρ΄ όλα αυτά, κατά τη διάρκεια της αλληλεπίδρασης η απόκριση της συμπεριφοράς και της προσοχής από τον άνθρωπο δεν είναι άμεση, καθώς μπορεί να υπάρχει καθυστερημένη επίδραση μετά από μια δράση του ρομπότ. Για τη μοντελοποίηση της διαδικασίας αυτής, επιλέγουμε μία δυναμικού τύπου συνάρτηση επιβράβευσης η οποία βασίζεται στην εικονική συμμετοχή E του ανθρώπου που λαμβάνει μέρος. Η συμμετοχή αυτή θα συσχετίζεται με την προσοχή που δίνει ο άνθρωπος στο ρομπότ και θα αναπαριστάται από ένα σήμα επιβράβευσης. Σε όλες τις προσομοιώσεις που θα ακολουθήσουν, η ποσοτικοποιημένη συμμετοχή θα ξεκινά από την τιμή 5, θα αυξάνεται σταδιακά έως την τιμή $E_M = 10$ όταν το ρομπότ πραγματοποιεί τις κατάλληλες δράσεις με τις κατάλληλες παραμέτρους, και θα πέφτει έως την τιμή $E_m = 0$ σε αντίθετη περίπτωση, όπως περιγράφεται παρακάτω.

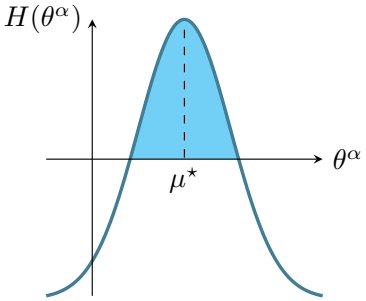
$$E_{t+1} = \begin{cases} E_t + \eta_1(E_M - E_t)H(\boldsymbol{\theta}_t^a), & \text{αν } a_t = a^\star \;\; H(\boldsymbol{\theta}_t^a) \geq 0 \\ E_t - \eta_2(E_m - E_t)H(\boldsymbol{\theta}_t^a), & \text{αν } a_t = a^\star \;\; H(\boldsymbol{\theta}_t^a) < 0 \\ E_t + \eta_2(E_m - E_t), & \text{αλλιώς} \end{cases} \tag{6.20}$$

όπου $\eta_1 = 0.1$ είναι ο ρυθμός αύξησης, και $\eta_2 = 0.05$ είναι ο ρυθμός μείωσης της συμμετοχής, ενώ $H(\boldsymbol{x})$ είναι η συνάρτηση ενεργοποίησης της συμμετοχής και ορίζεται ως,

$$H(\boldsymbol{x}) = 2\exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}^\star)^T(\boldsymbol{\Sigma}^\star)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}^\star)\right) - 1$$

όπου $a^\star$ είναι η βέλτιστη δράση, $\boldsymbol{\mu}^\star$ είναι το βέλτιστο διάνυσμα παραμέτρων $\boldsymbol{\theta}_\star^a$ της βέλτιστης δράσης και $\boldsymbol{\Sigma}^\star$ είναι ένας διαγώνιος πίνακας $\sigma^{\star 2}\boldsymbol{I}$ μεγέθους $m_{a^\star} \times m_{a^\star}$. Για να γίνει πιο αντιληπτό, οι τιμές των παραμέτρων για τις οποίες $H(\boldsymbol{x}) = 0$ ορίζουν το όρια μιας $m_{a^\star}$-διάστατης σφαίρας εντός του παραμετρικού χώρου, τέτοιας ώστε αν επιλεγεί η βέλτιστη δράση $a^\star$ και η παράμετροι της δράσης είναι εντός της σφαίρας, τότε η συμμετοχή θα αυξάνεται. Γενικά κάθε παράμετρος μπορεί να έχει τη δική της ανοχή, ωστόσο εδώ θα χρησιμοποιήσουμε ένα κοινό $\sigma^\star = 10$ ενώ το εύρος των τιμών θα είναι στο διάστημα $[-100, 100]$. Η εικόνα 6.5 απεικονίζει

τη συνάρτηση H στην περίπτωση όπου ο παραμετρικός χώρος της βέλτιστης δράσης είναι μονοδιάστατος.



Σχήμα 6.5: Συνάρτηση ενεργοποίησης της εικονικής συμμετοχής/προσοχής. Στη μονοδιάστατη περίπτωση, το ευθύγραμμο τμήμα στον χώρο παραμέτρων στο οποίο $H(\theta^a) > 0$, ορίζει το διάστημα ανοχής. Παράμετροι εντός του διαστήματος αυτού αυξάνουν την εικονική συμμετοχή/προσοχή (αν και εφόσον η διακριτή δράση a που επιλέχθηκε είναι επίσης η βέλτιστη). Στη γενική περίπτωση n διαστάσεων η αντίστοιχη περιοχή θα καθορίζεται από τα όρια ενός ελλειψοειδούς, ή υπερσφαίρας για κανονικοποιημένους παραμετρικούς χώρους.

Στις περιπτώσεις κατά τις οποίες μας ενδιαφέρουν δύο υποεργασίες, όπου για παράδειγμα η πρώτη μπορεί να αφορά την επίλυση ενός παζλ ενώ η δεύτερη την μεγιστοποίηση της ανθρώπινης συμμετοχής/προσοχής, μπορούμε να χρησιμοποιήσουμε μία υβριδική συνάρτηση επιβράβευσης, μορφοποιημένη ώστε να περιέχει μία συνιστώσα r^e, σχετική με τη μεγιστοποίηση της συμμετοχής, και μία συνιστώσα r^c, σχετικής με την επίλυση του παζλ.

Ακολουθώντας τη λογική από το [30], η επιβράβευση r^e για την επιβράβευση της αύξησης της προσοχής θα είναι

$$r^e_{t+1} = E_{t+1} + \lambda \Delta E_{t+1} \tag{6.21}$$

όπου $\lambda = 0.7$ είναι ένα βάρος επιλογής. Η παραπάνω εναλλακτική μορφοποίηση είναι χρήσιμη για την άμεση επιβράβευση δράσεων $(a, \boldsymbol{\theta}^a)$ κατά τις οποίες η συμμετοχή/προσοχή του ανθρώπου είναι μεν χαμηλή αλλά η δράση αυτή συμβάλλει στην άμεση αύξησή της. Η συνιστώσα r^c θα εξαρτάται γενικά από το πρόβλημα προς επίλυση. Στις περιπτώσεις όπου το ζητούμενο είναι η μεγιστοποίηση της προσοχής, τότε θα λαμβάνεται υπόψιν μόνο η r^e συνιστώσα. Για τις επόμενες προσομοιώσεις χρησιμοποιούμε τα σήματα $\bar{\delta}_V$ και $\bar{\delta}_Q$ για τη ρύθμιση της αντιστρόφου θερμοκρασίας β και της αβεβαιότητας των παραμέτρων των δράσεων, όπως αναφέρθηκε κατά την περιγραφή του αλγορίθμου.

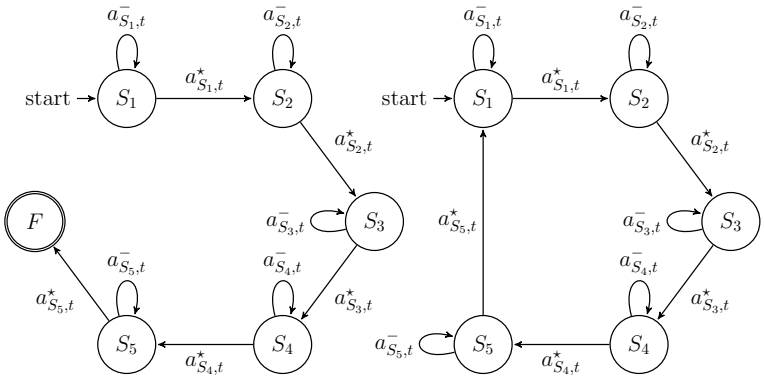
6.4.1 Αριθμητικές Προσομοιώσεις σε Περιβάλλον με 5 Μη-Στατικές Καταστάσεις

Επεκτείνοντας τις προσομοιώσεις που παρουσιάζονται στο [29], εδώ ελέγχουμε τις επιδόσεις του αλγορίθμου σε μία μαρκοβιανή διαδικασία λήψης αποφάσεων με παραμετροποιημένες δράσεις, σε έναν πεπερασμένο χώρο 5 εσωτερικών καταστάσεων και μίας τελικής επιπλέον κατάστασης (σχήμα 6.6), όπου ο σκοπός είναι η μεγιστοποίηση της συνάρτησης συμμετοχής.

Ο χώρος των δράσεων είναι $A = A_d \times A_p$, όπου $A_d = \{a_1, a_2, a_3, a_4, a_5, a_6\}$ και $A_p = [-100, 100]$. Με απλά λόγια υπάρχουν 6 διακριτές δράσεις με μία παράμετρο η κάθε μια. Υποθέτουμε ότι σε κάθε κατάσταση s_i σε κάθε χρονική στιγμή t, υπάρχει μόνο μία βέλτιστη δράση $a^\star_{s_i}$ η οποία οδηγεί σε αλλαγή κατάστασης στην επόμενη κατα σειρά. Αρχικά σε αυτό το σύνολο προβλημάτων η μετάβαση αυτή θα γίνεται ντετερμινιστικά (αργότερα θα παρουσιάσουμε και αποτελέσματα σε στοχαστικές μεταβάσεις), έτσι ώστε $\mathbb{P}[S_{t+1} = s_{i+1} | S_t = s_i, A_t = a^\star_{s_i}] = 1$, ανεξάρτητα από την επιλογή της τιμής της παραμέτρου $\theta^a_{s_i}$. Για όλες τις άλλες δράσεις $a \in A_d \backslash \{a^\star_{s_i}\}$ υποθέτουμε ότι η κατάσταση παραμένει αμετάβλητη. Κάθε βέλτιστη δράση $a^\star_{s_i}$ χαρακτηρίζεται και από μία βέλτιστη τιμή παραμέτρου $\mu^\star_{s_i}$ όπως φαίνεται στο σχήμα 6.5, ενώ η επιλογή παραμέτρου τέτοιας ώστε $H(\theta^a_{s_i}) > 0$ θα έχει ως αποτέλεσμα την αύξηση της εικονικής συμμετοχής και συνεπώς την αύξηση της επιστρεφόμενης επιβράβευσης r_t, όπως καθορίζεται από την εξίσωση 6.21. Αυτό θα συμβαίνει όταν

$$\mu^\star_{s_i} - \sigma^\star \sqrt{2ln2} < \theta^a_{s_i} < \mu^\star_{s_i} + \sigma^\star \sqrt{2ln2}$$

όπου το $\sigma^\star$ θα έχει την ίδια τιμή για όλες τις δράσεις σε όλες τις καταστάσεις. Αυτή η



Σχήμα 6.6: Μαρκοβιανή διαδικασία λήψης αποφάσεων 5 καταστάσεων. Αριστερά: MDP για μία απλή συνεδρία. Η βέλτιστη δράση $a^\star_{S_i,t}$ σε κάθε κατάσταση S_i οδηγεί στην αλλαγή κατάστασης στην επόμενη κατά σειρά. Όλες οι υπόλοιπες δράσεις $a^-_{S_i,t}$ δεν οδηγούν σε αλλαγή κατάστασης. Δεξιά: Αντίστοιχο MDP που χρησιμοποιήθηκε για την επαναληπτική διαδικασία αξιολόγησης.

ανισότητα περιγράφει το *διάστημα ανοχής* το οποίο θα έχει σταθερό εύρος (εφόσον υποθέτουμε σταθερό $\sigma^\star$) αλλά με μη στατικά όρια καθώς το περιβάλλον είναι δυναμικό, και η βέλτιστη παράμετρος $\mu^\star$ αλλά και η βέλτιστη διακριτή δράση $a^\star$ μπορεί να αλλάξουν. Με την ύπαρξη επιβράβευσης μετά από κάθε δράση (υψηλής ή χαμηλής), μπορεί κανείς να διαφωνήσει στο κατά πόσο επιδεικνύονται πραγματικά οι ικανότητες του αλγορίθμου στα πλαίσια της ενισχυτικής μάθησης, καθώς το πρόβλημα θα μπορούσε να προσεγγιστεί έχοντας ανεξάρτητα MABs, ένα σε κάθε κατάσταση. Παρ' όλα αυτά η μετάδοση των επιβραβεύσεων από δράσεις σε μελλοντικές καταστάσεις δεν θα ήταν ανιχνεύσιμες, ενώ στην προκειμένη περίπτωση κάτι τέτοιο είναι εφικτό.

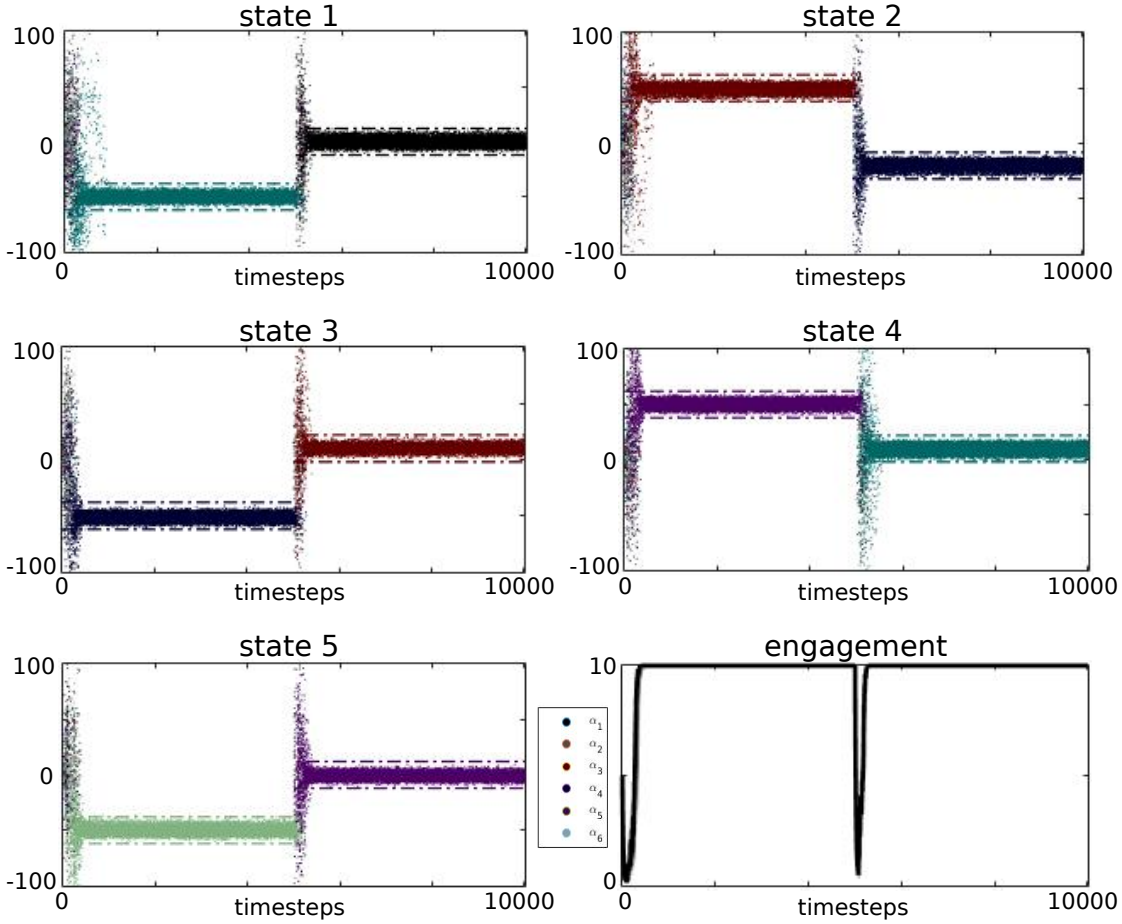
Αρχικά θα μελετηθούν προβλήματα όπου το περιβάλλον προς μελέτη θα υπόκειται σε καθολικές αλλαγές διακοπτόμενου τύπου ενώ στη συνέχεια οι αλλαγές θα είναι τοπικές (διακοπτόμενου και ομαλού τύπου). Σε κάθε περίπτωση η ρύθμιση των υπερπαραμέτρων έγινε όπως περιγράφηκε στην προηγούμενη παράγραφο.

Περιβάλλον με καθολικές αλλαγές διακοπτόμενου τύπου

Στη συνέχεια προσομοιώνουμε τον αλγόριθμο στο χρησιμοποιώντας το δεξί MDP του σχήματος 6.6 για 10000 χρονικές στιγμές, όπου και πάλι θα αναφερόμαστε σε αυτή τη διαδικασία ως υπερ-συνεδρία. Για τις χρονικές στιγμές $1 \leq t < 5000$ οι βέλτιστες διακριτές δράσεις θα είναι $\{a^\star_{s_1,t}, a^\star_{s_2,t}, a^\star_{s_3,t}, a^\star_{s_4,t}, a^\star_{s_5,t}\} = \{a_2, a_3, a_4, a_5, a_6\}$ και οι αντίστοιχες βέλτιστες τιμές παραμέτρων $\{\mu^\star_{s_1,t}, \mu^\star_{s_2,t}, \mu^\star_{s_3,t}, \mu^\star_{s_4,t}, \mu^\star_{s_5,t}\} = \{-50, 50, -50, 50, -50\}$. Για $t \geq 5000$ οι βέλτιστες διακριτές δράσεις θα είναι $\{a^\star_{s_1,t}, a^\star_{s_2,t}, a^\star_{s_3,t}, a^\star_{s_4,t}, a^\star_{s_5,t}\} = \{a_1, a_4, a_3, a_2, a_5\}$, ενώ οι βέλτιστες τιμές των αντίστοιχων παραμέτρων $\{\mu^\star_{s_1,t}, \mu^\star_{s_2,t}, \mu^\star_{s_3,t}, \mu^\star_{s_4,t}, \mu^\star_{s_5,t}\} = \{0, -10, 10, 10, 0\}$. Στο σχήμα 6.7 φαίνονται τα αποτελέσματα μετά από 50 υπερ-συνεδρίες. Η γραφικές παραστάσεις δείχνουν για κάθε κατάσταση από τις 5, τα ζεύγη δράσης-παραμέτρου που επιλέχθηκαν στις χρονικές στιγμές που η μηχανή μάθησης βρέθηκε στην κατάσταση αυτή. Το χρώμα των οριζόντιων διακεκομμένων γραμμών αναπαριστά το ποια ήταν η βέλτιστη δράση στην κατάσταση αυτή για την αντίστοιχη χρονική στιγμή, ενώ η περιοχή μεταξύ των διακεκομμένων γραμμών αντιστοιχεί στο διάστημα ανοχής, εντός του οποίου η συνάρτηση συμμετοχής είναι θετική. Η κάτω δεξιά γραφική παράσταση δείχνει την τιμή της μέσης συμμετοχής/προσοχής που επιτεύχθηκε από τις 50 υπερ-συνεδρίες, σε κάθε χρονική στιγμή.

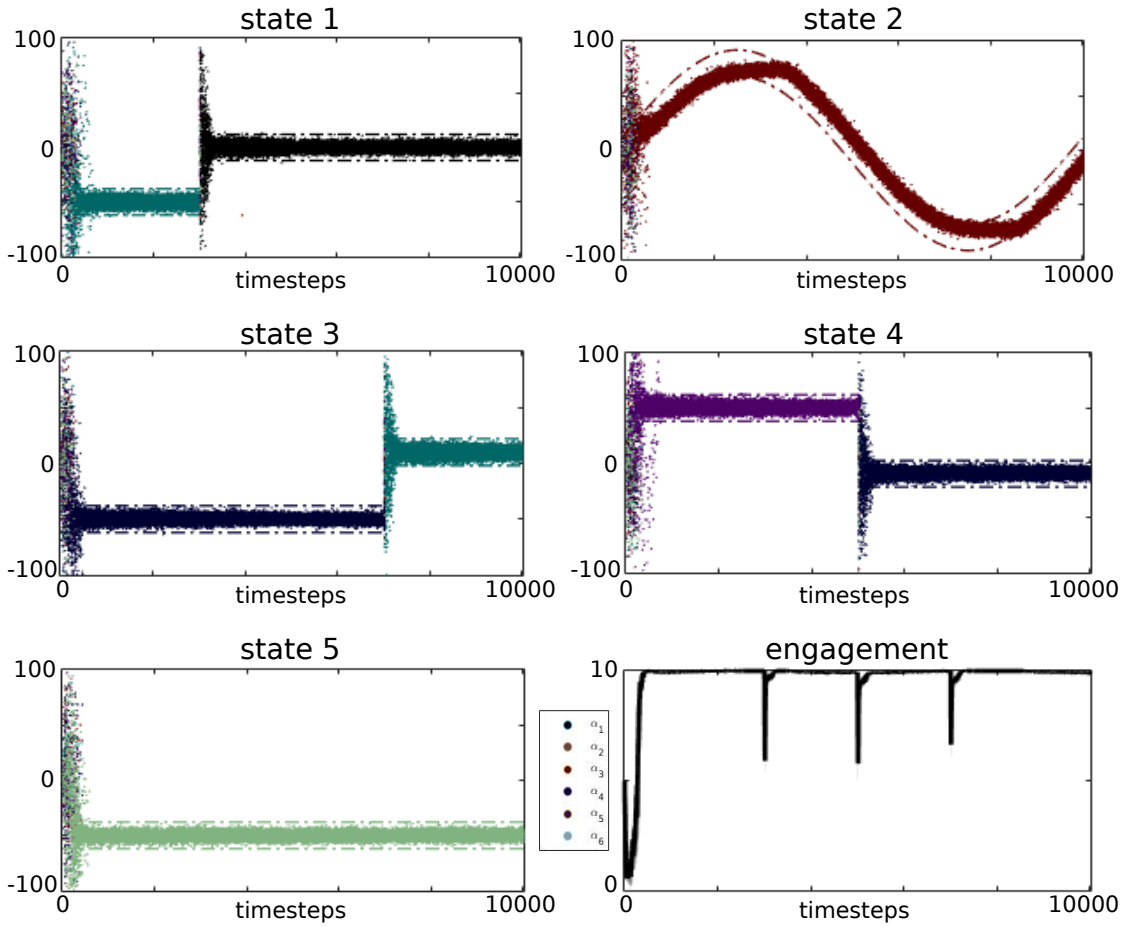
Η μηχανή μάθησης αρχικά εκτελεί εξερευνητικές δράσεις, κάτι το οποίο είναι εμφανές από το διάσπαρτο "νέφος δράσεων" από χρωματιστές κουκκίδες στο πρώτο χρονικό διάστημα, ενώ στη συνέχεια καταφέρνει να εκτιμήσει αρκετά καλά το βέλτιστο ζεύγος δράσης-παραμέτρου $(a^\star_{s_i}, \mu^\star_{s_i})$ σε όλες τις καταστάσεις s_i. Αυτό είναι εμφανές καθώς το χρώμα των κουκκίδων συμπίπτει με το χρώμα των διακεκομμένων γραμμών (σε αυτή την περίπτωση η διακριτή δράση είναι η βέλτιστη), καθώς επίσης οι κουκκίδες βρίσκονται εντός του διαστήματος που ορίζουν οι διακεκομμένες γραμμές (άρα η τιμή της παραμέτρου βρίσκεται εντός του διαστήματος ανοχής).

Η αβεβαιότητα των επιλεγομένων παραμέτρων, η οποία ρυθμίζεται δυναμικά με τη χρήση της συνάρτησης $\mathcal{F}_\sigma(\cdot)$, όπως περιγράφηκε στον μαθηματικό φορμαλισμό του αλγορίθμου, είναι κάτω φραγμένη από $\sigma_{\min} = 2$, κάτι το οποίο φαίνεται λόγω του κατακόρυφου εύρους του



Σχήμα 6.7: Αποτελέσματα σε μη στατικό περιβάλλον με καθολικές αλλαγές διακοπτόμενου τύπου. Οι επιλεγμένες δράσεις για 50 υπερ-συνεδρίες φαίνονται ως χρωματιστές κουκκίδες σε κάθε κατάσταση και κάθε χρονική στιγμή, όπου το κάθε χρώμα αντιστοιχεί σε μία διαφορετική δράση. Η τιμή της επιλεγμένης παραμέτρου αντιστοιχεί στην κατακόρυφη θέση του κάθε σημείου. Το χρώμα των διακεκομμένων γραμμών αντιστοιχεί στη βέλτιστη διακριτή δράση, ενώ το κατακόρυφο εύρος μεταξύ των διακεκομμένων γραμμών αντιστοιχεί στο διάστημα ανοχής. Η μέση εικονική συμμετοχή/παρακολούθηση από 50 υπερ-συνεδρίες για κάθε χρονική στιγμή φαίνεται κάτω δεξιά.

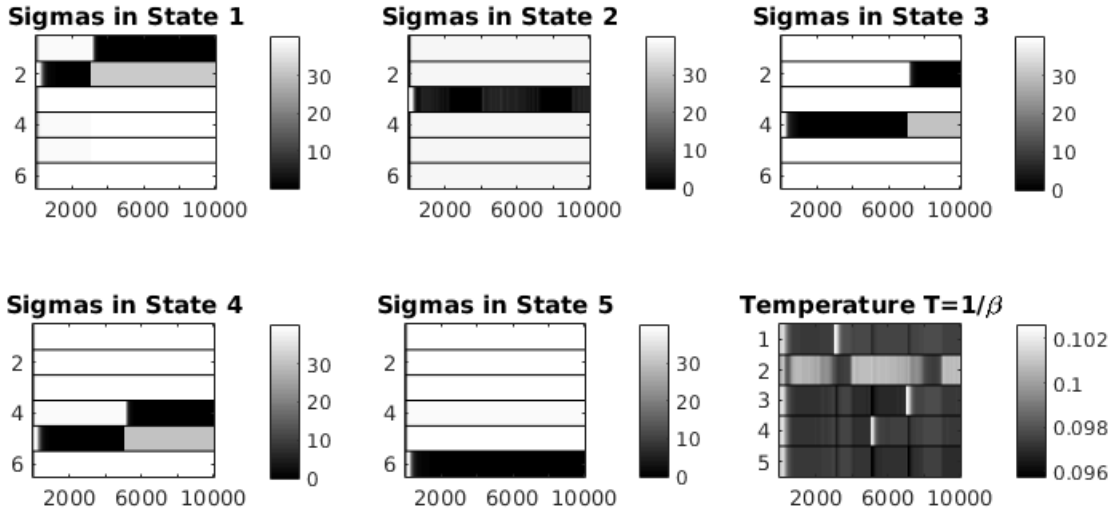
"νέφους δράσεων". Η εικονική συμμετοχή E_t αρχικά μειώνεται αλλά στη συνέχεια αυξάνεται ως το 10, επιτυγχάνοντας βέλτιστη τιμή πριν την αλλαγή του περιβάλλοντος. Ακριβώς μετά την αλλαγή, τα επίπεδα εξερευνητικών διακριτών δράσεων και οι αβεβαιότητες των τιμών των παραμέτρων τους αυξάνονται, βρίσκοντας στη συνέχεια τα νέα βέλτιστα ζεύγη $(a^\star, \mu^\star)$ ανά κατάσταση, και επιτυγχάνοντας τη βέλτιστη επίδοση μετά από περίπου 200 χρονικές στιγμές.



Σχήμα 6.8: Αποτελέσματα σε μη στατικό περιβάλλον με τοπικές και ομαλές αλλαγές. Σ΄ αυτή την περίπτωση κάθε κατάσταση μπορεί να αλλαζει σε χρονικές στιγμές ανεξάρτητα από τις υπόλοιπες, ενώ η βέλτιστη παράμετρος στην κατάσταση 2 αλλάζει με ημιτονοειδή τρόπο στην πάροδο του χρόνου. Η περιγραφή της διάταξης είναι παρόμοια με αυτή του σχήματος 6.7.

Περιβάλλον με τοπικές αλλαγές διακοπτόμενου και ομαλού τύπου

Στο δεύτερο αυτό πρόβλημα, η κάθε κατάσταση μπορεί να υπόκειται σε αλλαγές ανεξαρτήτως της δυναμικής των υπολοίπων. Όπως φαίνεται από την εικόνα 6.8 (όπου και πάλι οι διακεκομμένες γραμμές καθορίζουν τη βέλτιστη διακριτή δράση και το διάστημα ανοχής της παραμέτρου), στην κατάσταση s_1 υπάρχει μία διακοπτόμενη αλλαγή τη χρονική στιγμή $t = 3000$. Το βέλτιστο ζεύγος δράσης-παραμέτρου για $t < 3000$ είναι $(a_2, -50)$, ενώ αλλάζει σε $(a_1, 0)$ για $t \geq 3000$. Στην κατάσταση s_2 η βέλτιστη διακριτή δράση είναι η a_3 καθ΄ όλη τη διάρκεια της υπερ-συνεδρίας, ωστόσο η βέλτιστη τιμή της παραμέτρου έχει ημιτονοειδή εξάρτηση με τον χρόνο. Στην κατάσταση s_3 υπάρχει μία διακοπτόμενη αλλαγή τη χρονική στιγμή $t = 7000$, καθώς για $t < 7000$ το βέλτιστο ζεύγος δράσης-παραμέτρου είναι $(a_4, -50)$ ενώ για $t \geq 7000$ αλλάζει σε $(a_2, 0)$. Η κατάσταση s_4 είναι επίσης μη στατική, με βέλτιστο

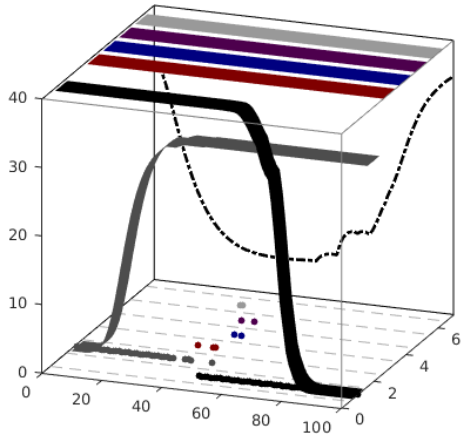


Σχήμα 6.9: Αβεβαιότητα σε μη στατικά περιβάλλοντα με τοπικές αλλαγές. Κάτω δεξιά: Απεικονίζεται η μέση θερμοκρασία $T(s) = 1/\beta(s)$ για κάθε μία από τις 5 καταστάσεις (γραμμές), σε κάθε χρονική στιγμή. Υπόλοιπες απεικονίσεις: Εμφάνιση της μέσης αβεβαιότητας $\sigma(s, a)$ για κάθε δράση (γραμμές) σε κάθε κατάσταση, για κάθε χρονική στιγμή.

ζεύγος $(a_5, 50)$ για $t < 5000$ και $(a_6, -10)$ για $t \geq 5000$. Η κατάσταση s_5 είναι στατική, με βέλτιστο ζεύγος $(a_6, -50)$.

Από τα αποτελέσματα της εικόνας 6.8 είναι εμφανές ότι τα βέλτιστα ζεύγη $(a^\star, \mu^\star)$ προσεγγίζονται σε αρκετά καλό βαθμό για κάθε χρονική στιγμή. Το "νέφος δράσεων" βρίσκεται εντός του διαστήματος ανοχής σε πολύ μεγάλο βαθμό, ενώ η εικονική συμμετοχή είναι πολύ κοντά στη μέγιστη δυνατή. Παρ' όλο που η βέλτιστη τιμή της παραμέτρου της δράσης a_3 στην κατάσταση s_2 αλλάζει διαρκώς, οι εξερευνητικές δράσεις φαίνεται να ακολουθούν δυναμικά κατά μέση τιμή τις αλλαγές αυτές, επιδεικνύοντας την ουσία και την ισχύ του αλγορίθμου. Επίσης είναι σημαντικό να παρατηρήσουμε ότι το ποσοστό των εξερευνητικών δράσεων στην κατάσταση s_5 δεν αυξάνεται, ακόμα και μετά την ύπαρξη διακοπτόμενων αλλαγών σε κάποια άλλη κατάσταση. Στο σχήμα 6.9(κάτω δεξιά) φαίνονται τα επίπεδα της μέσης θερμοκρασίας $T_t(s) = 1/\beta_t(s)$ σε όλες τις καταστάσεις και για όλες τις χρονικές στιγμές. Ως παράδειγμα, στην κατάσταση s_1, η απότομη μείωση των επιβραβεύσεων λόγω της αλλαγής του βέλτιστου ζεύγους δράσης-παραμέτρου, είχε ως αποτέλεσμα τη μείωση της αντιστρόφου θερμοκρασίας $\beta_t(s_1)$, η οποία φαίνεται στο σχήμα από την αύξηση της φωτεινότητας στην πρώτη γραμμή. Στην κατάσταση s_2 όπου η βέλτιστη τιμή της παραμέτρου αλλάζει με ομαλό τρόπο, παρατηρούμε ότι η εικονική θερμοκρασία παραμένει υψηλή, εκτός από τα μικρά διαστήματα στα οποία η μονοτονία του ημίτονου αλλάζει, και η υπάρχουσα εκτίμηση της βέλτιστης παραμέτρου βρίσκεται εκτός του διαστήματος ανοχής. Παράλληλα, στο ίδιο σχήμα φαίνονται και οι αβεβαιότητες για τις παραμέτρους της καθε δράσης σε κάθε κατάσταση και κάθε χρονική στιγμή. Κατά την ύπαρξη μίας διακοπτόμενης αλλαγής, παρατηρούμε την αύξηση της αβεβαι-

ότητας για την τιμή της παραμέτρου της προηγούμενης βέλτιστης δράσης, και την μείωση της αβεβαιότητας για την παράμετρο της νέας.



Σχήμα 6.10: Συνολική εικόνα προσαρμοστικής πολιτικής σε διακοπτόμενη αλλαγή. Στη συγκεκριμένη εικόνα φαίνεται ένα χρονικό διάστημα λήψης αποφάσεων από τον αλγόριθμο, σε μία από τις καταστάσεις, κατά το οποίο υπήρχε μία διακοπτόμενη αλλαγή του βέλτιστου ζεύγους δράσης-παραμέτρου. Ο x-άξονας είναι οι χρονικές στιγμές στις οποίες η μηχανή μάθησης βρέθηκε σε αυτή την κατάσταση και στον y-άξονα φαίνονται οι 6 διαφορετικές δράσεις προς επιλογή. Στον z-άξονα οι 6 διαφορετικές "ταινίες" αφορούν τις αβεβαιότητες $\sigma(s, a)$ για την παράμετρο κάθε δράσης, ενώ στο πίσω μέρος (διακεκομμένη γραμμή) απεικονίζεται η αντίστροφη θερμοκρασία $\beta(s)$.

Για να έχουμε μία ακόμα καλύτερη εικόνα του τρόπου με τον οποίο λειτουργεί η ενεργός εξερεύνηση ανά κατάσταση, στην εικόνα 6.10 απεικονίζονται τα χαρακτηριστικά που μας ενδιαφέρουν σε κάποιο χρονικό διάστημα μίας συνεδρίας κατά το οποίο συμβαίνει μία διακοπτόμενη αλλαγή. Στο συγκεκριμένο παράδειγμα ο αλγόριθμος δεν χρησιμοποιεί τις βέλτιστες τιμές των παραμέτρων ώστε να είναι πιο εύκολη η παρατήρηση της δυναμικής αλλαγής των χαρακτηριστικών αυτών (σε άλλη περίπτωση η προσαρμοστικότητα μπορεί να είναι πολύ υψηλή και να μην βοηθάει στην παρατήρηση των χαρακτηριστικών). Μετά από τη διακοπτόμενη αλλαγή στην κατάσταση s_1, η αβεβαιότητα για τη βέλτιστη τιμή της παραμέτρου αρχίζει να αυξάνεται ενώ το ίδιο συμβαίνει και για την αβεβαιότητα του είδους της βέλτιστης διακριτής δράσης, κάτι που φαίνεται από την μείωση της αντιστρόφου θερμοκρασίας $\beta(s)$ στην κατάσταση αυτή. Μόλις η αβεβαιότητα αυξηθεί αισθητά, η μηχανή μάθησης αρχίζει με μεγαλύτερη πιθανότητα να επιλέγει κάποια άλλη διακριτή δράση (κάτι το οποίο αναπαριστάται από τις διαφορετικού χρώματος κουκκίδες στο κάτω μέρος του σχήματος). Όταν η βέλτιστη δράση επιλεχθεί και η αντίστοιχη παράμετρος είναι εντός του νέου διαστήματος ανοχής, τότε η επίδοση αρχίζει να βελτιώνεται, η αβεβαιότητα για την παράμετρο αυτής της νέας δράσης αρχίζει να μειώνεται και η αντίστροφη θερμοκρασία αυξάνεται, ξεκινώντας να αυξάνεται παράλληλα η αξιοποιητική πολιτική των δράσεων.

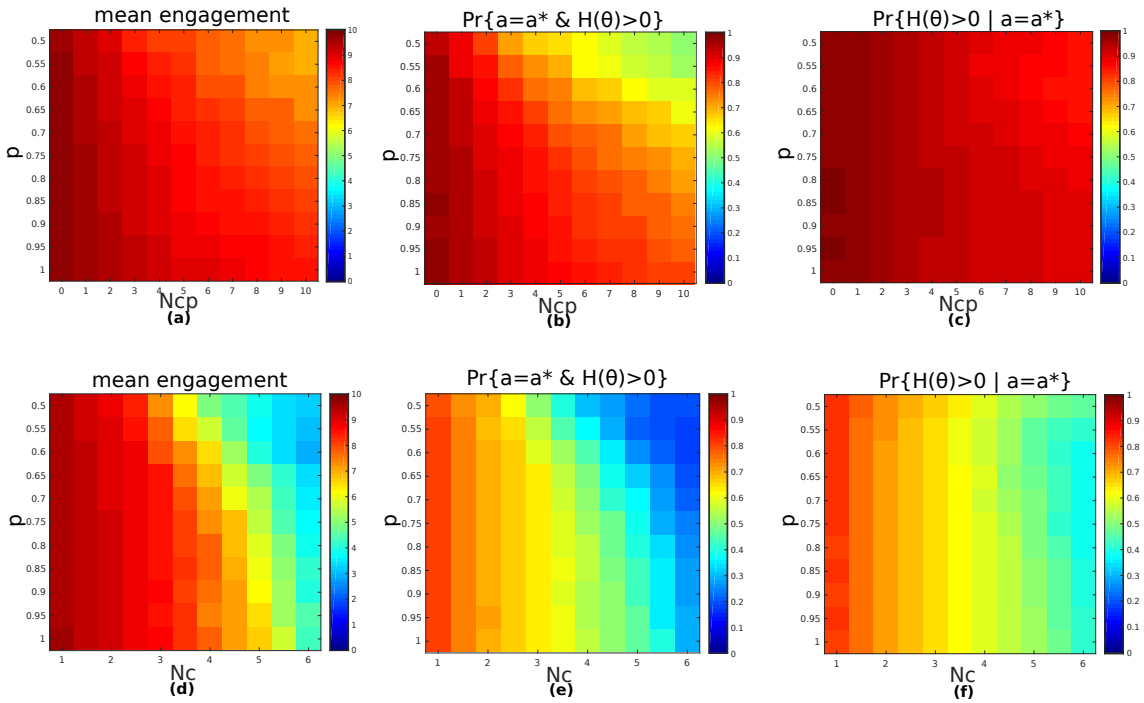
Στο παράρτημα Α′, παρουσιάζονται τα αντίστοιχα αποτελέσματα της προηγούμενης έκδοσης του αλγορίθμου, στην οποία τα επίπεδα εξερεύνησης και αβεβαιότητας ήταν κοινά μεταξύ δράσεων και καταστάσεων.

Ευρωστία ως προς τη στοχαστικότητα και τη μεταβλητότητα

Για την επιπλέον αξιολόγηση των επιδόσεων του αλγορίθμου σε περιβάλλοντα με στοχαστικές μεταβάσεις και την αξιολόγηση της προσαρμοστικότητάς του σε σχέση με τα επίπεδα μεταβλητότητας, προσομοιώνουμε τον αλγόριθμο σε μία στοχαστική έκδοση του περιβάλλοντος του σχήματος 6.6, τροποποιώντας την πιθανότητα των ορθών μεταβάσεων από 0.5 έως 1. Συγκεκριμένα, υποθέτουμε ότι όταν επιλέγεται η βέλτιστη δράση, τότε η μετάβαση στην επόμενη κατάσταση πραγματοποιείται με πιθανότητα p ενώ με πιθανότητα $1-p$ η κατάσταση δεν αλλάζει. Αντίστοιχα, όταν η δράση που επιλέγεται δεν είναι η βέλτιστη, τότε η κατάσταση αλλάζει με πιθανότητα $1-p$ ενώ παραμένει η ίδια με πιθανότητα p.

Καθώς μελετούμε την επίδοση για διαφορετικές τιμές του p απο 0.5 έως 1 με βήμα 0.05, τροποποιούμε παράλληλα την μεταβλητότητα του περιβάλλοντος με δύο διαφορετικούς τρόπους, με διακοπτόμενες αλλαγές και αλλαγές ομαλού τύπου. Αρχικά, αλλάζουμε το βέλτιστο ζεύγος δράσης-παραμέτρου με κυκλική εναλλαγή. Αν n είναι ένας δείκτης αρχικοποιημένος στο 0, τέτοιος ώστε σε κάθε αλλαγή του περιβάλλοντος $n \leftarrow n+1$, τότε η βέλτιστη δράση σε κάθε κατάσταση s_i θα δίνεται από τη σχέση $a^\star_{s_i} = a_j$, όπου $j = \mod(i+n-1, 6)+1$, ενώ η βέλτιστη τιμή της παραμέτρου της δράσης αυτής θα είναι $\mu^\star_{s_i} = (-1)^{n+i} \times 20$. Επίσης αν N_{cp} συμβολίζει τον αριθμό των αλλαγών του περιβάλλοντος (ομοιόμορφα κατανεμημένων στο χρονικό εύρος από 1-10000), τότε τροποποιούμε την τιμή του από 0 έως 10 και συνολικά αξιολογούμε τις επιδόσεις του αλγορίθμου για 20 υπερ-συνεδρίες σε κάθε πιθανό ζευγάρι (p, N_{cp}). Υπολογίζουμε τη μέση εικονική συμμετοχή που επιτυγχάνεται και υπολογίζουμε αριθμητικά τις πιθανότητες $\mathbb{P}[a = a^\star \& H(\theta^a) > 0]$ και $\mathbb{P}[H(\theta^a) > 0 | a = a^\star]$, όπως φαίνεται στο σχήμα 6.11(πάνω). Επίσης, σταθεροποιούμε τα βέλτιστα ζεύγη δράσεων-παραμέτρων με κάποια τυχαία αρχικοποίηση ανά κατάσταση και τροποποιούμε τη βέλτιστη τιμή των παραμέτρων με ημιτονοειδή εξάρτηση στον χρόνο. Συμβολίζοντας με N_c τον αριθμό των κύκλων του ημίτονου από την αρχή έως το τέλος κάθε υπερ-συνεδρίας, αξιολογούμε τις επιδόσεις του αλγορίθμου για όλα τα διαφορετικά ζευγάρια (p, N_c) που προκύπτουν τροποποιώντας τον αριθμό των κύκλων N_c από $N_c = 1$ έως $N_c = 6$ με βήμα 0.5. Τα αποτελέσματα φαίνονται στο σχήμα 6.11.

Για να έχουμε ένα κοινό σημείο αναφοράς, επισημαίνουμε ότι η μέση επίδοση που παρουσιάστηκε στο περιβάλλον των καθολικών αλλαγών του σχήματος 6.7, αντιστοιχεί στην τιμή για $p = 1$ και $N_{cp} = 1$ στο σχήμα 6.11a. Στα διακοπτόμενου τύπου περιβάλλοντα με χαμηλή μεταβλητότητα η μείωση του p δε φαίνεται να επιφέρει σημαντική μείωση της απόδοσης. Παρ′ όλα αυτά δεν συμβαίνει το ίδιο όταν το N_{cp} αυξάνεται. Στα περιβάλλοντα με υψηλή μεταβλητότητα και υψηλή στοχαστικότητα η απόδοση είναι η χαμηλότερη, όπως άλλωστε ήταν αναμενόμενο, πετυχαίνοντας ωστόσο μέση απόδοση γύρω στο 75% στη χειρότερη περίπτωση. Στα περιβάλλοντα ομαλών μεταβάσεων, οι επιδόσεις ήταν αρκετά υψηλές για N_c από 1-3, χωρίς η επίδραση



Σχήμα 6.11: Εκτίμηση ευρωστίας σε στοχαστικότητα και μεταβλητότητα. (a,b,c): Επιδόσεις για διαφορετικά επίπεδα στοχαστικότητας p και μεταβλητότητας τύπου N_{cp}, όπου σε κάθε κατάσταση αλλάζει με διακοπτόμενου τύπου αλλαγή η βέλτιστη διακριτή δράση και η παράμετρος αυτής. (d,e,f): Επιδόσεις για διαφορετικά επίπεδα στοχαστικότητας p και μεταβλητότητας τύπου N_c, όπου οι βέλτιστες παράμετροι όλων των δράσεων αλλάζουν με ημιτονοειδή τρόπο.

της στοχαστικότητας να είναι εμφανής. Η επιδόσεις αρχίζουν να μειώνονται δραστικά μετά από αυτό το σημείο με χειρότερη μέση εικονική συμμετοχή γύρω στο 35%.

6.4.2 Αλληλεπίδραση Ρομπότ-Παιδιού με Επίλυση Υποπροβλήματος

Στη συνέχεια υλοποιούμε μία εικονική προσομοίωση αλληλεπίδρασης ρομπότ-παιδιού σε ένα σενάριο όπου 3 χρωματιστοί κύβοι διαφορετικού μεγέθους ο κάθε ένας (μικρού μεγέθους, μεσαίου μεγέθους, μεγάλου μεγέθους) είναι τοποθετημένοι με τυχαία σειρά σε 3 πιθανές θέσεις (αριστερά, κέντρο, δεξιά), στην επιφάνεια ενός τραπεζιού. Το ρομπότ πρέπει να μάθει να κατασκευάζει στην κέντρική θέση έναν πύργο τοποθετώντας τον έναν κύβο πάνω στον άλλο, ακολουθώντας τους κανόνες του παζλ «ο πύργος του Ανόι[11]» και έχοντας ως μόνη ανάδραση μία μικρή επιβράβευση κάθε φορά που κατασκευάζει τον πύργο με επιτυχία (επι-

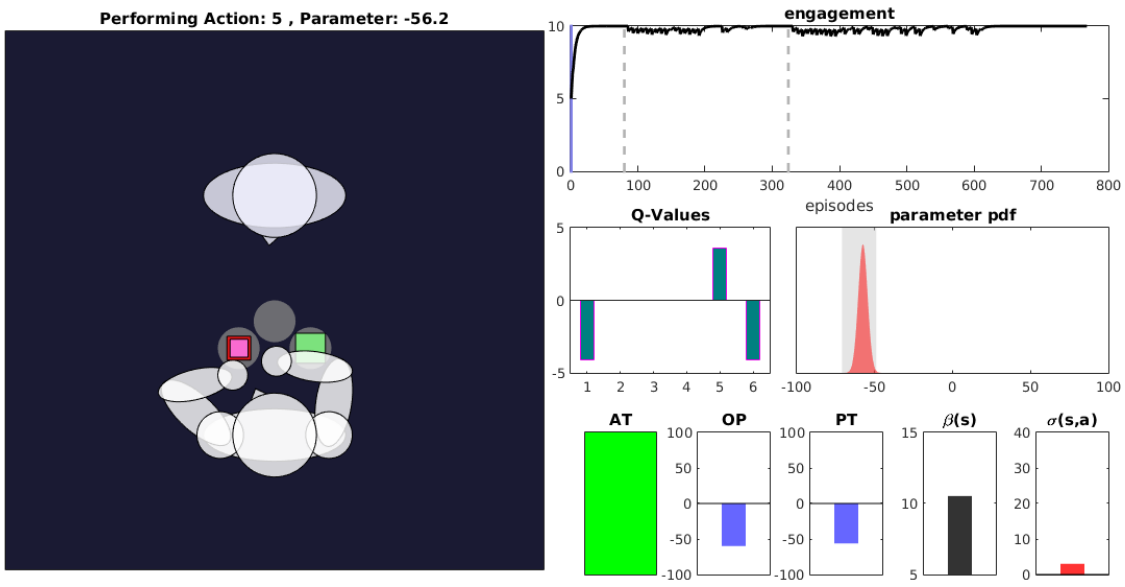
[11] Αν και το παζλ συνήθως περιγράφεται με τόρους και στύλους, εδώ υποθέτουμε ότι δεν επιτρέπεται ένας μεγαλύτερου μεγέθους κύβος να βρίσκεται πάνω σε κάποιον μικρότερου μεγέθους. Σε ένα πραγματικό περιβάλλον, εάν μία απαγορευμένη κίνηση εκτελεστεί από το ρομπότ, υποθέτουμε ότι μπορεί να υπάρξει ένα σήμα διακοπής ή λεκτική εντολή από τον άνθρωπο ώστε να αντιστρέψει την κίνηση, οπότε η κατάσταση θα παραμείνει αμετάβλητη. Αυτός είναι και ο τρόπος με τον οποίο έχει υλοποιηθεί εσωτερικά η προσομοίωση. Οι κινήσεις μπορούν να επιλεχθούν, αλλά η κατάσταση δεν αλλάζει.

βράβευση μη κοινωνικού τύπου) καθώς και ένα οπτικό ερέθισμα από το παιδί (επιβράβευση κοινωνικού τύπου). Το ρομπότ είναι εφοδιασμένο με την ικανότητα να πιάνει και να τοποθετεί αντικείμενα από τις τρεις προκαθορισμένες θέσεις, πιάνοντας κάθε φορά το αντικείμενο που βρίσκεται στην κορυφή, και τοποθετώντας το στην κορυφή της στοίβας στο επιλεγμένο σημείο προορισμού. Μπορεί επίσης να τροποποιεί χαρακτηριστικά που έχουν να κάνουν με την κίνηση, όπως ταχύτητα, ομαλότητα, εκφραστικότητα. Για παράδειγμα μπορεί διαμέσου κατάλληλης επιλογής παραμέτρων να αλλάζει τη συχνότητα με την οποία ανταλλάζει ματιές με το παιδί πριν πιάσει έναν κύβο, ώστε να μεγιστοποιήσει την κοινή προσοχή προς την επίτευξη του στόχου. Μπορεί επίσης να επιλέγει τροχιές με μεγαλύτερη ή μικρότερη καμπυλότητα, ή να μεταβάλει την ταχύτητα της κίνησης των άκρων του δυναμικά ώστε να δίνει την αίσθηση μίας προσεκτικά ρυθμισμένης κίνησης για αποφυγή λαθών. Στις προσομοιώσεις που ακολουθούν θα θεωρήσουμε μία παράμετρο, η οποία θα επηρεάζει την ταχύτητα της κίνησης.

Θεωρούμε ότι το πρόβλημα μπορεί να περιγραφεί από 27 καταστάσεις, όπου η κάθε κατάσταση εμπεριέχει πληροφορία για την τρέχουσα διάταξη των κύβων στις 3 θέσεις. Για παράδειγμα, η κατάσταση s_1 αντιστοιχεί στη διάταξη κατά την οποία ο μικρός κύβος βρίσκεται στην αριστερότερη θέση, ο μεσαίου μεγέθους κύβος βρίσκεται στην κεντρική θέση, ενώ ο μεγαλύτερος κύβος βρίσκεται στη δεξιότερη θέση. Το διάνυσμα κατάστασης $\boldsymbol{\phi}(s)$ θα είναι τύπου one-hot, έτσι ώστε $\phi_i(s) = \mathbb{I}\{s = s_i\}$.

Υπάρχουν 6 διαθέσιμες διακριτές δράσεις συνολικά. Με τη δράση a_1 το ρομπότ προσπαθεί να πιάσει τον κύβο που βρίσκεται στην κορυφή της αριστερότερης θέσης (ελέγχοντας πρώτα αν υπάρχει διαθέσιμος κύβος) και να τον τοποθετήσει στην κορυφή της στοίβας που βρίσκεται στο κέντρο. Εάν χρησιμοποιήσουμε τον συμβολισμό "LC" γι' αυτή την κίνηση (από την αγγλική απόδοση Left to Center), τότε ο χώρος διακριτών δράσεων A_d θα είναι $\{a_1, a_2, a_3, a_4, a_5, a_6\} = \{LC, LR, CL, CR, RL, RC\}$, χωρίς ωστόσο να επιφέρουν όλες οι δράσεις αλλαγή κατάστασης, καθώς είτε δεν επιτρέπονται (οπότε η κίνηση θα αντιστραφεί μετά το τέλος της), είτε δεν θα υπάρχει διαθέσιμος κύβος στην αφετηρία (οπότε η δράση δεν θα εκτελεστεί). Οι παράμετροι μπορούν να πάρουν τιμές στο διάστημα [-100, 100] στο οποίο γίνεται και η αναζήτηση, όπως και στην περίπτωση των MDPs, ενώ έπειτα μετασχηματίζονται κατάλληλα προς τους επενεργητές του ρομπότ.

Θεωρούμε ότι υπάρχουν 6 διαφορετικές οικογένειες δράσεων, όπου η κάθε οικογένεια προϋποθέτει διαφορετικά επίπεδα δεξιότητας της κίνησης του ρομπότ. Οι δράσεις *τύπου 1* θα είναι αυτές κατά τις οποίες το ρομπότ πιάνει έναν μοναχικό κύβο (κύβος ο οποίος δεν βρίσκεται πάνω σε άλλον), και τον τοποθετεί σε μία άδεια θέση (θέση στην οποία δεν υπάρχει άλλος κύβος). Αυτού του είδους οι δράσεις δεν περιέχουν κάποια επικινδυνότητα, είτε κατά την αρπαγή του κύβου είτε κατά την τοποθέτηση, και θεωρούμε ότι μπορούν να πραγματοποιηθούν με σχετικά υψηλή ταχύτητα. Θα χρησιμοποιήσουμε τον συμβολισμό "$1-1$" για τις δράσεις αυτές. Υποθέτουμε επίσης ότι κάθε τύπος δράσης θα χαρακτηρίζεται από μία βέλτιστη παράμετρο με την οποία η κίνηση του ρομπότ θα μοιάζει πιο ρεαλιστική. Χωρίς βλάβη της γενικότητας, εδώ έχουμε μοντελοποιήσει το πρόβλημα με τέτοιο τρόπο ώστε η συμμετοχή/προσοχή του παιδιού να αυξάνεται όταν οι κινήσεις του ρομπότ έχουν ρεαλιστική ταχύτητα ανάλογα με τον τύπο της δράσης. Έτσι, όταν μία κίνηση προϋποθέτει υψηλού βαθμού δεξιότητα, το



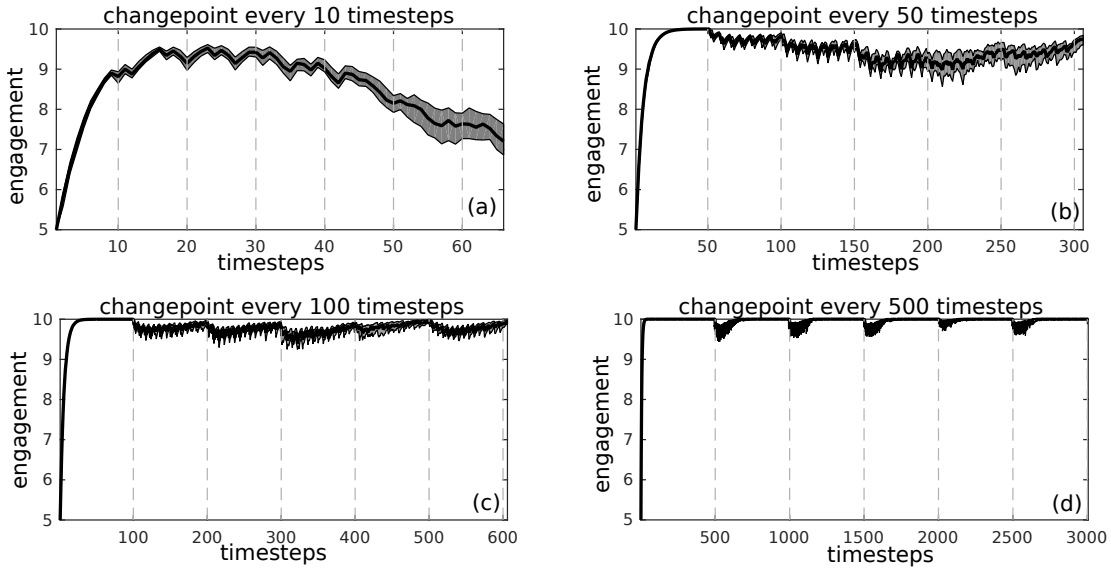
Σχήμα 6.12: Στιγμιότυπο εικονικού περιβάλλοντος αλληλεπίδρασης ρομπότ-παιδιού. Αριστερά: Στιγμιότυπο κατά το οποίο το ρομπότ έχει ήδη πιάσει τον μικρό ροζ κύβο πάνω από τον μεγάλο πράσινο κύβο, και τον έχει μόλις τοποθετήσει πάνω στον μεσαίου μεγέθους κόκκινο κύβο. Θεωρούμε ότι η κατάσταση δεν έχει αλλάξει. AT: Εκτελούμενη δράση (πράσινο χρώμα αν είναι η βέλτιστη, κόκκινο χρώμα αν δεν είναι η βέλτιστη). OP: Βέλτιστη τιμή παραμέτρου. PT: Τιμή παραμέτρου που επιλέχθηκε. $\beta(a)$: Αντίστροφη θερμοκρασία στην κατάσταση αυτή. $\sigma(s,a)$: Τυπική απόκλιση για την Gaussian εξερεύνηση, σχετιζόμενη με την αβεβαιότητα της επιλεγόμενης τιμής της παραμέτρου. Q-Values: Οι αξίες των δράσεων (ακριβώς πριν την εκτέλεση της συγκεκριμένης δράσης). Μέση δεξιά: Συνάρτηση πυκνότητας πιθανότητας από την οποία δειγματοληπτήθηκε η παράμετρος της δράσης. Η σκιασμένη περιοχή αφορά το διάστημα ανοχής. Πάνω: Εικονική συμμετοχή/προσοχή μετά από κάθε δράση. Οι διακεκομμένες γραμμές συμβολίζουν χρονικά σημεία στα οποία η βέλτιστη τιμή κάποιας παραμέτρου άλλαξε.

ρομπότ θα πρέπει να δίνει την αίσθηση ότι την εκτελεί προσεκτικά, αυξάνοντας παράλληλα την προσοχή του παιδιού. Έτσι, όλες οι δράσεις "$1-1$" θα χαρακτηρίζονται από μία κοινή βέλτιστη παράμετρο $\mu_1^\star$. Οι δράσεις *τύπου 2* θα είναι αυτές κατά τις οποίες το ρομπότ πιάνει τον κύβο που βρίσκεται στην κορυφή μίας στοίβας 2 κύβων, και τον τοποθετεί σε μία άδεια θέση. Συμβολίζουμε αυτές τις δράσεις με "$2-1$", για τις οποίες η βέλτιστη παράμετρος θα είναι $\mu_2^\star$. Ακολουθώντας την ίδια λογική θα έχουμε συνολικά 6 διαφορετικούς τύπους δράσεων, $\{$"$1-1$","$2-1$","$3-1$","$1-2$","$2-2$","$1-3$"$\}$, ταξινομημένες σε αύξουσα σειρά δεξιότητας της κίνησης. Όπως ήδη αναφέραμε, η συμμετοχή/προσοχή του παιδιού θα συσχετίζεται με τη "φυσικότητα" της κίνησης, ωστόσο υποθέτουμε ότι καθώς επιλύεται το παζλ, θα υπάρχει και μία σταδιακή αύξηση της βέλτιστης ταχύτητας δράσεων (είτε λόγω εξοικείωσης είτε λόγω ανυπομονησίας), και συνεπώς θα υπάρχει και σταδιακή αύξηση των βέλτιστων παραμέτρων

για κάθε τύπο δράσης ξεχωριστά. Το ρομπότ θα πρέπει λοιπόν να επιδείξει χαρακτηριστικά προσαρμοστικότητας.

Στο σχήμα 6.12 φαίνεται ένα στιγμιότυπο του περιβάλλοντος που αναπτύχθηκε για τους σκοπούς της προσομοίωσης. Στο στιγμιότυπο αυτό, το ρομπότ έχει μόλις τοποθετήσει τον μικρό ροζ κύβο (ο οποίος προηγουμένως βρισκόταν πάνω στον μεγάλο πράσινο κύβο) πάνω στον μεσαίου μεγέθους κόκκινο κύβο στην αριστερότερη θέση. Η εικονική συμμετοχή/παρακολούθηση του παιδιού εκφράζεται από τη γωνία του βλέμματος του παιδιού σε σχέση με το σημείο ενδιαφέροντος. Στο περιβάλλον αυτό, η γωνία του βλέμματος δειγματοληπτείται από μια biomodal κατανομή πιθανότητας, ως το άθροισμα δύο επιμέρους Gaussian κατανομών με μέσους ω^+ και ω^- και σταθερή μικρή διασπορά η κάθε μία, όπου η γωνία $(\omega^+ + \omega^-)/2$ είναι η γωνία προς το σημείο ενδιαφέροντος ενώ το μέγεθος $|\omega^+ - \omega^-|$ είναι αντιστρόφως ανάλογο της συμμετοχής/παρακολούθησης. Έτσι, στις περιπτώσεις που το ενδιαφέρον του παιδιού είναι χαμηλό, τότε κοιτάζει δεξιά αριστερά από το σημείο ενδιαφέροντος με μεγάλη πιθανότητα. Το σημείο ενδιαφέροντος σε αυτή την υλοποίηση θεωρούμε ότι είναι η θέση της κινούμενης παλάμης του ρομπότ. Είναι προφανές ότι θα μπορούσαμε να μοντελοποιήσουμε όλα τα παραπάνω με πολύ μεγαλύτερη λεπτομέρεια, όμως κάτι τέτοιο είναι εκτός του σκοπού της εργασίας. Πίσω στο σχήμα 6.12, το πράσινου χρώματος παραλληλόγραμμο πλαίσιο με περιγραφή AT (Action Taken) περιγράφει αν η δράση που εκτελέστηκε είναι η βέλτιστη (οπότε το πλαίσιο θα έχει πράσινο χρώμα όπως στο στιγμιότυπο), ή όχι (οπότε το πλαίσιο θα έχει κόκκινο χρώμα). Στα δεξιά του πλαισίου AT φαίνεται η βέλτιστη τιμή της παραμέτρου στο πλαίσιο OP (Optimal Parameter), ενώ στο επόμενο πλαίσιο PT (Parameter Taken) φαίνεται η τιμή της παραμέτρου που επιλέχθηκε κατά την τελευταία δράση. Από τα τρία αυτά πλαίσια (AT-OP-PT) μπορούμε να δούμε ότι πράγματι το ζεύγος δράσης-παραμέτρου που επιλέχθηκε ήταν το βέλτιστο (όσον αφορά τη διακριτή δράση μπορούμε να το επιβεβαιώσουμε και με επισκόπηση του προβλήματος). Στο πλαίσιο $\beta(s)$ φαίνονται τα επίπεδα της αντιστρόφου θερμοκρασίας. Αν και η τιμή αυτή προσαρμόζεται μέσω του ίδιου του αλγορίθμου, μία τιμή κοντά στο 11 αντιστοιχεί σε υψηλά επίπεδα αξιοποιητικών δράσεων. Στο πλαίσιο $\sigma(s, a)$ φαίνεται η αβεβαιότητα με την οποία επιλέχθηκε η παράμετρος, η οποία έχει πάρει την ελάχιστη δυνατή τιμή $\sigma_{\min} = 2$. Ακριβώς από πάνω φαίνεται η κατανομή $p(\theta^a|a, s)$ για τον συγκεκριμένο τύπο δράσης, καθώς και το διάστημα ανοχής. Αριστερά φαίνονται οι εκτιμώμενες αξίες των δράσεων. Η δράση a_5 που αντιστοιχεί σε κίνηση RL, δηλαδή μετακίνηση κύβου από την δεξιότερη θέση στην αριστερότερη θέση, είναι αυτή που έχει την υψηλότερη τιμή και ήταν αυτή που επιλέχθηκε μέσω της soft-max. Η γραφική παράσταση στην υψηλότερη θέση απεικονίζει την εικονική συμμετοχή/παρακολούθηση του παιδιού σε κάθε χρονική στιγμή, ενώ οι διακεκομμένες γραμμές αντιστοιχούν σε σημεία στα οποία άλλαξε η βέλτιστη παράμετρος κάποιου τύπου δράσεων. Μία πιο πλήρης περιγραφή των χαρακτηριστικών της πλατφόρμας υπάρχει επίσης στο συμπληρωματικό υλικό του [60].

Για την περαιτέρω αξιολόγηση της προσαρμοστικότητας και των ορίων αυτής, διεξήγαμε μία πληθώρα από προσομοιώσεις, αλλάζοντας τη συχνότητα εμφάνισης αλλαγής της βέλτιστης παραμέτρου κάποιου τύπου δράσης. Συμβολίζοντας με T_{cp} τον αριθμό των κινήσεων



Σχήμα 6.13: Ευρωστία της αλληλεπίδρασης για διαφορετικά επίπεδα μεταβλητότητας με απεικόνιση της μέσης εικονικής συμμετοχής/παρακολούθησης μετά από 50 υπερσυνεδρίες. (a): Αλλαγή της βέλτιστης τιμής κάποιας παραμέτρου κάθε 10 κινήσεις. (b): Αλλαγή της βέλτιστης τιμής κάποιας παραμέτρου κάθε 50 κινήσεις. (c): Αλλαγή της βέλτιστης τιμής κάποιας παραμέτρου κάθε 100 κινήσεις. (d): Αλλαγή της βέλτιστης τιμής κάποιας παραμέτρου κάθε 500 κινήσεις.

μετά από τις οποίες πραγματοποιείται η αλλαγή[12], επαναλάβαμε 20 υπερ-συνεδρίες για $T_{cp} \in \{10, 50, 100, 500\}$. Στο σχήμα 6.13 απεικονίζεται η μέση εικονική συμμετοχή/παρακολούθηση που επιτεύχθηκε σε κάθε βήμα, για κάθε διαφορετική τιμή του T_{cp}. Για $T_{cp} = 10$ φαίνεται ότι ο αλγόριθμος δεν μπόρεσε να ακολουθήσει τις αλλαγές και οι επιδόσεις ήταν πολύ χαμηλές. Για $T_{cp} = 50$ φαίνεται ότι υπήρχε μέτρια προσαρμοστικότητα. Για $T_{cp} = 100$ επιτεύχθηκε μεγιστοποίηση της εικονικής συμμετοχής/παρακολούθησης ακριβώς πριν από κάθε αλλαγή, ενώ για $T_{cp} = 500$ όπου τα στατικά διαστήματα ήταν μεγάλα υπήρχε αρκετός χρόνος για μεγιστοποίηση και σταθεροποίηση της επίδοσης.

6.5 Συνολική Αξιολόγηση και Συζήτηση

Σε αυτό το κεφάλαιο παρουσιάσαμε την ανάπτυξη ενός νέου αλγορίθμου ενισχυτικής μάθησης σε παραμετροποιημένους χώρους δράσεων και χρήση ενεργούς εξειδικευμένης εξερεύνησης ανά κατάσταση σε μη στάσιμα περιβάλλοντα, επεκτείνοντας σε μεγάλο βαθμό τις ιδέες που παρουσιάστηκαν στα [29, 30]. Η επέκταση αυτή επιτρέπει τη διαχείριση σεναρίων με πολλές καταστάσεις, όπου η κάθε κατάσταση μπορεί να ακολουθεί διαφορετική δυναμική. Τα αποτε-

[12]Σε κάθε αλλαγή η βέλτιστη παράμετρος κάποιας οικογένειας δράσεων (εδώ επιλέχθηκαν οι δράσεις "$1-2$") άλλαζε κατά 10 μονάδες. Αρχικά η αλλαγή είχε θετικό πρόσημο ενώ κάθε φορά που έφτανε στα όρια του παραμετρικού χώρου το πρόσημο της αλλαγής αντιστρεφόταν.

λέσματα επιδεικνύουν το δυναμικό του αλγορίθμου και η πιθανή συνεισφορά του στον χώρο της μηχανικής μάθησης μπορεί να έχει διαφορετικές προεκτάσεις. Για παράδειγμα, εάν η διάσπαση ενός σύνθετου προβλήματος σε ένα σύνολο μικρότερων γενικευμένων και παραμετροποιημένων προβλημάτων είναι εφικτή, τότε ο αλγόριθμος αυτός ενδέχεται να επιτυγχάνει πολύ καλές επιδόσεις, όχι μόνο σε στάσιμες περιπτώσεις αλλά ακόμα και όταν υπάρχει άγνωστη και μη ανιχνεύσιμη χρονομεταβλητή εξάρτηση της λύσης. Απο θεωρητική σκοπιά, προτείνεται μία νέα μέθοδος για εξειδικευμένη εξερεύνηση ανά κατάσταση επεκτείνοντας τις μεθόδους ενεργούς αναζήτησης [49, 4, 43, 7]. Αναφορικά με τις επιπτώσεις που μπορεί να έχει στον τομέα της ρομποτικής μάθησης, φαίνεται πλέον ότι η επιλογή παραμετροποιημένων χώρων δράσεων (αντί της χρήσης του συνεχή χώρου μεγάλων διαστάσεων και αβέβαιη εύρεση επίλυσης), μπορεί να είναι πολύ ελκυστική. Από την σκοπιά των εφαρμογών αλληλεπίδρασης ανθρώπουρομπότ, ο αλγόριθμος φαίνεται να επιτυγχάνει καλύτερα αποτελέσματα προσαρμοστικότητας σε μη λεκτικά σήματα από τις προϋπάρχουσες έρευνες [17, 11, 47, 25, 38, 51], ενώ προτείνει μία εμπειρικά καλή και εφικτή λύση για τις περιπτώσεις όπου παρουσιάζονται διακυμάνσεις της συμμετοχής/προσοχής κατά την αλληλεπίδραση [25, 51, 1]. Τέλος, σε εκπαιδευτικές εφαρμογές στις οποίες ένα μικρό ανθρωποειδές ρομπότ έχει τον ρόλο υποβοήθησης ενός δάσκαλουεκπαιδευτικού, ο προτεινόμενος αλγόριθμος μπορεί να αποτελέσει μία καινοτόμα προσέγγιση στη βελτίωση της αλληλεπίδραση παιδιών υπό το φάσμα του αυτισμού (ASD) σε εκπαιδευτικά παιχνίδια, και υποβοήθηση για ανάπτυξη των κοινωνικών τους δεξιοτήτων [26, 48, 5] και του ενδιαφέροντος για νέες αναζητήσεις [11].

Κεφάλαιο 7

Επίλογος και Νέες Κατευθύνσεις

7.1 Γενική Αξιολόγηση

Στην παρούσα εργασία έγινε αρχικά μία περιγραφή των βασικότερων ιδεών και κινήτρων για την εφαρμογή του πλαισίου της ενισχυτικής μάθησης σε ρομποτικές εφαρμογές, με κύριο αντίκτυπο σε κοινωνικές προεκτάσεις μέσω της αλληλεπίδρασης ρομπότ-παιδιού. Το γενικό κίνητρο της εργασίας αφορά ρομπότ υποβοήθησης, ενισχυμένα με την ικανότητα δυναμικής προσαρμοστικότητας κατά την αναγνώριση μη-λεκτικών χαρακτηριστικών επικοινωνίας. Ως παράδειγμα αναφέραμε ένα εκπαιδευτικό περιβάλλον στο οποίο ένα μικρό ανθρωποειδές ρομπότ μπορεί να υποβοηθήσει τον άνθρωπο-εκπαιδευτικό ώστε να ενισχύσει το ενδιαφέρον παιδιών υπό το φάσμα του αυτισμού (ASD) σε εκπαιδευτικά παιχνίδια, να τα βοηθήσει για την ανάπτυξη κοινωνικών δεξιοτήτων και την ενίσχυση των κινήτρων τους. Το παράδειγμα αυτό καθορίζεται από το πλαίσιο του προγράμματος Ευρωπαϊκής Ένωσης BabyRobot (H2020-ICT-24-2015-6878310), και ήταν η βασική πηγή έμπνευσης για την ανάπτυξη των αλγορίθμων και των πειραμάτων που παρουσιάστηκαν. Στη συγκεκριμένη εργασία βέβαια δεν παρουσιάστηκε η εφαρμογή μεθόδων σε πραγματικά περιβάλλοντα, καθώς ο σκοπός της ήταν η ανάπτυξη νέων προσαρμοστικών αλγορίθμων στο χαμηλότερο επίπεδο σχεδίασης, και η αξιολόγηση των επιδόσεών τους σε μια ποικιλία από μη-στατικά προβλήματα, μίας αλλά και περισσότερων καταστάσεων. Όπως αναφέραμε στο κεφάλαιο 1, σημειώσαμε δύο βασικές συνεισφορές σε επίπεδο ανάπτυξης αλγορίθμων:

• την ανάπτυξη ενός υβριδικού προσαρμοστικού αλγορίθμου για λήψη αποφάσεων υπό αβεβαιότητα σε δυναμικά περιβάλλοντα μίας κατάστασης.

Πράγματι, αρχικά παρουσιάστηκε το πρόβλημα λήψης αποφάσεων υπό αβεβαιότητα σε μία κατάσταση, με κατάλληλη και περιεκτική περιγραφή των ήδη υπάρχοντων αλγορίθμων, αλλά και των βασικότερων προκλήσεων που μπορεί να παρουσιαστούν. Οι λόγοι που μας οδήγησαν σε αυτού του είδους τις αναζητήσεις, ήταν η προσπάθεια για μελέτη και ανάπτυξη εγγενών χαρακτηριστικών προσαρμοστικότητας. Συνδυάζοντας βιολογικά εμπνευσμένη μετα-μάθηση και έναν αλγόριθμο Bayesian εκτίμησης παραμέτρων, αναπτύξαμε ένα νέο αλγόριθμο (τον οποίο

ονομάσαμε MLB-KF). Οι εμπειρικές του επιδόσεις φαίνεται να ξεπερνούν (σαν μία μέση εικόνα) αυτές σύγχρονων προσαρμοστικών αλγορίθμων.

• την ανάπτυξη ενός προσαρμοστικού αλγορίθμου ενισχυτικής μάθησης με εξειδικευμένη εξερεύνηση ανά κατάσταση για παραμετροποιημένους χώρους δράσεων.

Πράγματι, στη συνέχεια παρουσιάστηκε το πλαίσιο της ενισχυτικής μάθησης και των μεθόδων που μπορούν να χρησιμοποιηθούν σε γνωστής και άγνωστης δομής Μαρκοβιανές διαδικασίες λήψης αποφάσεων, ενώ αναπτύχθηκε ένας νέος αλγόριθμος ενισχυτικής μάθησης με ενεργή εξερεύνηση ανά κατάσταση σε περιβάλλοντα με παραμετροποιήσιμες διακριτές δράσεις. Ο αλγόριθμος αυτός επέκτεινε την προηγούμενη υλοποίηση [30] και βελτίωσε τις επιδόσεις σε μη στατικά περιβάλλοντα όπου η δυναμική εξέλιξη κάθε κατάστασης είναι ανεξάρτητη των υπολοίπων. Αν και δεν υπήρχαν συγκριτικά αποτελέσματα, οι εμπειρικές επιδόσεις του αλγορίθμου φαίνονται πολλά υποσχόμενες. Το πλαίσιο Μαρκοβιανών διαδικασιών αποφάσεων με παραμετροποιημένες διακριτές δράσεις είναι νέο, και συνεπώς δεν υπάρχουν προσαρμοστικοί αλγόριθμοι ώστε να γίνει η κατάλληλη σύγκριση. Ένα σκεπτικό ήταν να τροποποιηθούν οι υπάρχοντες αλγόριθμοι [23], [41], κάτι που ωστόσο θα ξεπερνούσε τα πλαίσια της εργασίας, αλλά ενδεχομένως βρίσκεται στα πλάνα νέας έρευνας.

7.2 Ιδέες για Νέες Κατευθύνσεις Μελλοντικής Έρευνας

Οι ιδέες για νέες κατευθύνσεις μπορεί να ποικίλουν σε ειδικές και γενικές. Οι ειδικές αφορούν την εκ νέου βελτίωση ή τροποποίηση των αλγορίθμων που παρουσιάστηκαν, ενώ οι γενικές αφορούν τις νέες ιδέες για έρευνα με διαφορετικές προσεγγίσεις απο διαφορετική οπτική. Παρακάτω παραθέτουμε ένα μείγμα από πιθανές μελλοντικές επεκτάσεις.

Νέες Κατευθύνσεις σε Μία Κατάσταση

• Αρχικά είναι εμφανές ότι δεν υπάρχει μαθηματική ανάλυση σχετικά με την ασυμπτωτική συμπεριφορά του αλγορίθμου MLB-KF ως προς την αθροιστική μεταμέλεια. Αν και οι εμπειρικές επιδόσεις φαίνονται πολλά υποσχόμενες, οι προσδοκίες για την ύπαρξη ασυμπτωτικής μεταμέλειας κοντά στη βέλτιστη $O(\sqrt{T})$ είναι χαμηλές. Μία επέκταση λοιπόν είναι η αναλυτική αναζήτηση αυτής της ασυμπτωτικής συμπεριφοράς.

• Η χρήση της soft-max δίνει μία ιεραρχική στοχαστική ισορροπία στις εξερευνητικές δράσεις, όμως η μεταβολή του β από το 0 στο ∞ δημιουργεί αριθμητικά προβλήματα. Μία ιδέα είναι η χρήση προσαρμοστικής ϵ-greedy πολιτικής.

• Η χρήση Bayesian εκτίμησης με Βήτα κατανομές αντί κανονικών κατανομών και η αντίστοιχη υλοποίηση φίλτρων Kalman για μη στατικά περιβάλλοντα. Αν και έγιναν κάποιες δοκιμές και τα πειραματικά αποτελέσματα δεν ήταν εξίσου καλά με την υπάρχουσα υλοποίηση, θα πρέπει να γίνει ένας πιο προσεκτικός σχεδιασμός καθώς η συγκεκριμένη προσέγγιση έρχεται

σε συμφωνία με τη δειγματοληψία Thompson για στατικά περιβάλλοντα.

Νέες Κατευθύνσεις σε Περισσότερες Καταστάσεις

• Η χρήση καθολικής αβεβαιότητας για τις βέλτιστες παραμέτρους $\boldsymbol{\theta}_a^\star$ κάθε δράσης μπορεί να μην επαρκεί. Μία επέκταση είναι να χρησιμοποιήσουμε διαγώνιο πίνακα $\boldsymbol{\Sigma}_s^a$ με διαφορετική αβεβαιότητα ανά παράμετρο.

• Ο πίνακας $\boldsymbol{\Sigma}_s^a$ θα μπορούσε να μην είναι διαγώνιος. Κάτι τέτοιο μπορεί μάλιστα να είναι και το πιο σύνηθες σενάριο σε πραγματικές εφαρμογές. Μία βασική ιδέα για πιθανή μελλοντική έρευνα, είναι η ανανέωση της Gaussian εκτίμησης των παραμέτρων με χρήση φίλτρων Kalman, ενώνοντας έτσι τις ιδέες της έρευνάς μας σε μία κατάσταση με αυτές σε ενισχυτική μάθηση.

• Αν και η χρήση eligibility traces με γραμμική προσέγγιση συναρτήσεων δεν φαίνεται να έχει πολύ καλά αποτελέσματα [56], η υλοποίηση με μία διαφορετική βάση αλγορίθμου όπως Sarsa(λ) ή Q(λ) θα είχε ενδιαφέρον.

• Επέκταση σε πιο βαθιές αρχιτεκτονικές DQN. Κάτι τέτοιο μάλιστα ίσως έχει την προοπτική συνένωσης των μεθόδων που αναπτύξαμε με την υπάρχουσα δουλειά των Hausknecht και Stone [23].

Οι παραπάνω ιδέες αποτελούν ένα πολύ μικρό υποσύνολο όσων μπορεί να σκεφτεί κανείς, ειδικά σε ένα νέο αλλά διαρκώς εξελισσόμενο και γεμάτο ενδιαφέρον επιστημονικό πεδίο.

Bibliography

[1] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani. Evaluating the engagement with social robots. *International Journal of Social Robotics*, 7(4):465–478, 2015.

[2] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[4] A. Baranes and P.-Y. Oudeyer. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1):49–73, 2013.

[5] T. Belpaeme, P. E. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu, R. Looije, M. Neerincx, Y. Demiris, R. Ros-Espinoza, A. Beck, L. Cañamero, A. Hiolle, M. Lewis, I. Baroni, M. Nalin, P. Cosi, G. Paci, F. Tesser, G. Sommavilla, and R. Humbert. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, 1(2):33–53, 2012.

[6] F. Benureau and P.-Y. Oudeyer. Diversity-driven selection of exploration strategies in multi-armed bandits. In *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 135–142. IEEE, 2015.

[7] F. Benureau and P.-Y. Oudeyer. Behavioral diversity generation in autonomous exploration through reuse of past experience. *Frontiers in Robotics and AI*, 8, 2016.

[8] M. Brezzi and T. L. Lai. Incomplete learning from endogenous data in dynamic allocation. *Econometrica*, 68(6):1511–1516, 2000.

[9] G. Burtini, J. Loeppky, and R. Lawrence. A survey of online experiment design with the stochastic multi-armed bandit. *arXiv preprint arXiv:1510.00757*, 2015.

[10] H. Chen and F. Allgower. A quasi-infinite horizon nonlinear model predictive control scheme with guaranteed stability. *Automatica*, 34(10):1205–1217, 1998.

[11] K.-Y. Chin, Z.-W. Hong, and Y.-L. Chen. Impact of using an educational robot-based learning system on students? motivation in elementary education. *IEEE Transactions on learning technologies*, 7(4):333–345, 2014.

[12] N. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci*, 8(12):1704–1711, 2005.

[13] N. D. Daw and K. Doya. The computational neurobiology of learning and reward. *Current opinion in neurobiology*, 16(2):199–204, 2006.

[14] N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, and R. J. Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879, 2006.

[15] K. Doya. Metalearning and neuromodulation. *Neural Netw*, 15(4-6):495–506, 2002.

[16] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[17] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003.

[18] M. J. Frank, B. B. Doll, J. Oas-Terpstra, and F. Moreno. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature neuroscience*, 12(8):1062–1068, 2009.

[19] A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

[20] J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[21] O.-C. Granmo and S. Berg. *Solving Non-Stationary Bandit Problems by Random Sampling from Sibling Kalman Filters*, pages 199–208. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[22] C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag. Multi-armed bandit, dynamic environments and meta-bandits. 2006.

[23] M. Hausknecht and P. Stone. Deep reinforcement learning in parameterized action space. In *International Conference on Learning Representations (ICLR 2016)*. 2016.

[24] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[25] S. Ivaldi, S. Lefort, J. Peters, M. Chetouani, J. Provasi, and E. Zibetti. Towards engagement models that consider individual factors in hri: on the relation of extroversion and negative attitude towards robots to gaze and speech during a human-robot assembly task. *International Journal of Social Robotics*, 9(1):63–86, 2017.

[26] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction*, 19(1):61–84, 2004.

[27] M. Khamassi, P. Enel, P. Dominey, and E. Procyk. Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. *Progress in Brain Research*, 202:441–464, 2013.

[28] M. Khamassi, S. Lallée, P. Enel, E. Procyk, and P. Dominey. Robot cognitive control with a neurophysiologically inspired reinforcement learning model. *Frontiers in Neurorobotics*, 5:1, 2011.

[29] M. Khamassi, G. Velentzas, T. Tsitsimis, and C. Tzafestas. Active exploration and parameterized reinforcement learning applied to a simulated human-robot interaction task. In *IEEE Robotic Computing 2017 Conference*, pages 28–35, Taipei, Taiwan, 2017.

[30] M. Khamassi, G. Velentzas, T. Tsitsimis, and C. Tzafestas. Robot fast adaptation to changes in human engagement during simulated dynamic social interaction with active exploration in parameterized reinforcement learning. *IEEE Transactions on Cognitive and Developmental Systems*, 2018.

[31] J. Kober, J. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, pages 1238–1274, 2013.

[32] J. Kober and J. Peters. Policy search for motor primitives in robotics. *Machine Learning*, 84:171–203, 2011.

[33] L. Kocsis and C. Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.

[34] L. Kocsis and C. Szepesvári. Discounted ucb. In *2nd PASCAL Challenges Workshop*, pages 784–791, 2006.

[35] D. E. Koulouriotis and A. Xanthopoulos. Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems. *Applied Mathematics and Computation*, 196(2):913–922, 2008.

[36] G. Kreisselmeier and B. Anderson. Robust model reference adaptive control. *IEEE Transactions on Automatic Control*, 31(2):127–133, 1986.

[37] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[38] S. Lemaignan, M. Warnier, E. Sisbot, A. Clodic, and R. Alami. Artificial cognition for social human-robot interaction: An implementation. *Artificial Intelligence*, 247:45–69, 2017.

[39] S. Mahadevan and M. Maggioni. Proto-value functions: A laplacian framework for learning representation and control in markov decision processes. *Journal of Machine Learning Research*, 8(Oct):2169–2231, 2007.

[40] O.-A. Maillard, R. Munos, and G. Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 497–514, 2011.

[41] W. Masson and G. Konidaris. Reinforcement learning with parameterized actions. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*. 2016.

[42] J. C. Mellor. *Decision Making Using Thompson Sampling*. PhD thesis, University of Manchester, 2014.

[43] C. Moulin-Frier and P. Oudeyer. Exploration strategies in developmental robotics: a unified probabilistic framework. In *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–6. IEEE, 2013.

[44] J. Niño-Mora. A (2/3) n 3 fast-pivoting algorithm for the gittins index and optimal stopping of a markov chain. *INFORMS Journal on Computing*, 19(4):596–606, 2007.

[45] P. Parks. Liapunov redesign of model reference adaptive control systems. *IEEE Transactions on Automatic Control*, 11(3):362–367, 1966.

[46] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, pages 784–791. ACM, 2008.

[47] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner. Recognizing engagement in human-robot interaction. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, pages 375–382. IEEE, 2010.

[48] B. Robins, K. Dautenhahn, R. Te Boekhorst, and A. Billard. Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Universal Access in the Information Society*, 4(2):105–120, 2005.

[49] J. Schmidhuber. Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, 18(2):173–187, 2006.

[50] N. Schweighofer and K. Doya. Meta-learning in reinforcement learning. *Neural Networks*, 16(1):5–9, 2003.

[51] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164, 2005.

[52] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[53] A. Sokolov, J. Kreutzer, C. Lo, and S. Riezler. Learning structured predictors from bandit feedback for interactive nlp. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1610–1620, 2016.

[54] R. L. Solomon and J. D. Corbit. An opponent-process theory of motivation: I. temporal dynamics of affect. *Psychological review*, 81(2):119, 1974.

[55] F. Stulp and O. Sigaud. Robot skill learning: From reinforcement learning to evolution strategies. *Paladyn Journal of Behavioral Robotics*, 4(1):49–61, 2013.

[56] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[57] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[58] M. van der Meer, Z. Kurth-Nelson, and A. D. Redish. Information processing in decision-making systems. *The Neuroscientist*, 18(4):342–359, 2012.

[59] H. van Hasselt and M. Wiering. Reinforcement learning in continuous action spaces. In *IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 272–279. 2007.

[60] G. Velentzas, T. Tsitsimis, I. Rañó, C. Tzafestas, and M. Khamassi. Adaptive reinforcement learning with active state-specific exploration for engagement maximization during simulated child-robot interaction. *Paladyn, Journal of Behavioral Robotics*, 9(1):235–253.

[61] G. Velentzas, C. Tzafestas, and M. Khamassi. Bio-inspired meta-learning for active exploration during non-stationary multi-armed bandit tasks. In *IEEE Intelligent Systems Conference 2017*, London, UK, 2017.

[62] G. Velentzas, C. Tzafestas, and M. Khamassi. Bridging computational neuroscience and machine learning on non-stationary multi-armed bandits. *bioRxiv*, page 117598, 2017.

[63] J. Vermorel and M. Mohri. Multi-armed bandit algorithms and empirical evaluation. In *European conference on machine learning*, pages 437–448. Springer, 2005.

[64] J. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

[65] C. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Παράρτημα Α′

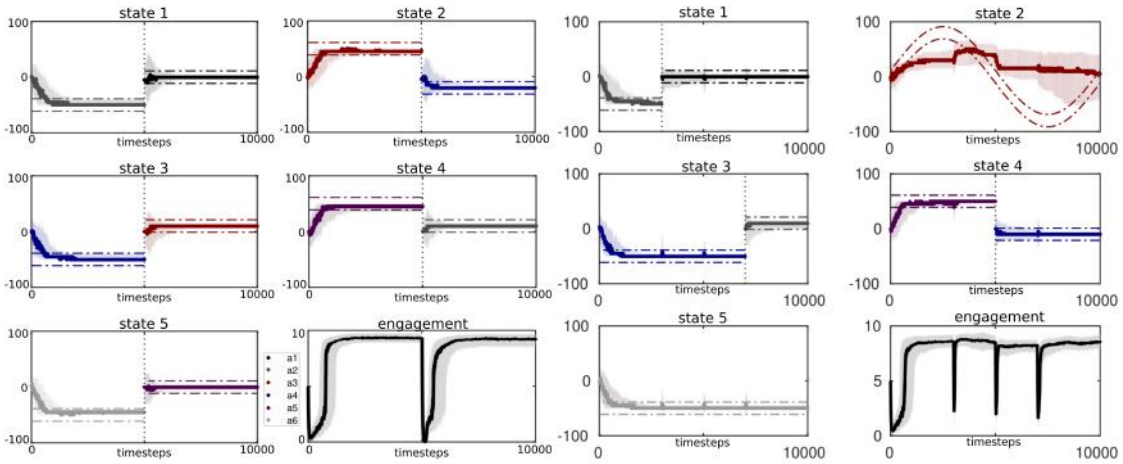
Σύγκριση με παλαιότερες υλοποιήσεις

Α′.0.1 Βελτίωση των επιδόσεων της παλαιότερης υλοποίησης

Παρακάτω επισυνάπτουμε τα αποτελέσματα της προηγούμενης έκδοσης [30] για τα μη στάσιμα περιβάλλοντα 5 καταστάσεων που περιγράφηκαν στο κεφάλαιο 6.

Περιβάλλον με καθολικές διακοπτόμενες αλλαγές.

Όπως φαίνεται στο σχήμα Α′.1 (αριστερή ομάδα γραφικών παραστάσεων 3×2), η οποία αφορά το πρόβλημα σε μη στατικό περιβάλλον με καθολικού τύπου αλλαγές, η μέγιστη εικονική συμμετοχή/παρακολούθηση που επιτυγχάνεται δεν είναι η βέλτιστη δυνατή. Επίσης μετά το σημείο αλλαγής η μέγιστη τιμή δεν φτάνει στα ίδια επίπεδα και η διασπορά στο σύνολο των 50 υπερ-συνεδριών είναι μεγαλύτερη από ότι στο πρώτο διάστημα από 1-5000. Ο



Σχήμα Α′.1: Αριστερά: Επίδοση στο αντίστοιχο περιβάλλον καθολικών διακοπτόμενων αλλαγών. Δεξιά: Επίδοση στο αντίστοιχο περιβάλλον τοπικών και ομαλών αλλαγών.

κύριος λόγος είναι η ύπαρξη ενός καθολικού μεγέθους που περιγράφει την αβεβαιότητα των διακριτών δράσεων αλλά και των επιλεγόμενων παραμέτρων τους. Έτσι, στις περιπτώσεις που οι περισσότερες δράσεις είναι οι βέλτιστες και κάποια δράση δεν είναι, τότε η συνολική απόδοση θα τείνει να μειώσει την εξερεύνηση, και η υπο-βέλτιστη δράση κινδυνεύει να μην προσαρμοστεί. Αυτό είναι κατά κύριο λόγο εμφανές στις καταστάσεις 5 και 4, όπου η μέση τιμή των παραμέτρων είναι κοντά στο άνω και κάτω όριο του διαστήματος ανοχής αντίστοιχα. Η συνολική επίδοση από τις άλλες δράσεις ωστόσο μειώνει την αβεβαιότητα, και συνεπώς ο αλγόριθμος λειτουργεί υποβέλτιστα. Ο νέος αλγόριθμος δεν πάσχει από αυτά τα προβλήματα και η εικονική συμμετοχή/προσοχή μεγιστοποιείται.

Περιβάλλον με τοπικές και ομαλού τύπου αλλαγές.

Στο δεξί μπλοκ 3×2 γραφικών του σχήματος Α΄.1 φαίνονται τα αποτελέσματα της πρώτης έκδοσης και υλοποίησης του αλγορίθμου. Όπως είναι εμφανές από το αποτέλεσμα ο αλγόριθμος λειτουργεί και πάλι υποβέλτιστα. Το χαρακτηριστικό που παρατηρούμε σε όλες τις καταστάσεις, είναι ότι η ύπαρξη αλλαγής σε κάποια άλλη κατάσταση επιφέρει αύξηση της εξερεύνησης. Αυτό είναι λογικό καθώς η αντίστροφη θερμοκρασία αλλά και η αβεβαιότητα των παραμέτρων είναι κοινή για όλες τις καταστάσεις και όλες τις δράσεις. Χαρακτηριστικό είναι ότι στην κατάσταση 2 ο αλγόριθμος αποτυγχάνει πλήρως να ακολουθήσει τη δυναμική αλλαγή του διαστήματος ανοχής.

Από τα πειραματικά αποτελέσματα είναι εμφανές ότι η νέα έκδοση του αλγορίθμου, εκτός από τα οφέλη που μπορεί να αντιληφθεί κανείς λόγω της πιο γενικής περιγραφής με χρήση διανυσμάτων κατάστασης και γραμμική προσέγγιση των περισσότερων μεγεθών μέσω αυτών, βελτιώνει τα αισθητά τα αποτελέσματα της προηγούμενης υλοποίησης.